Terminologie & Ontologie : Théories et Applications

Actes de la conférence

TOTh 2019

Le Bourget du Lac – 6 & 7 juin 2019

Les ouvrages TOTh précédents sont disponibles sur le site du Comptoir des Presses d'Universités (www.lcdpu.fr) ou auprès de : contact@toth.condillac.org

Éditeur : Presses Universitaires Savoie Mont Blanc
27 rue Marcoz
BP 1104
73011 CHAMBÉRY CEDEX
www.univ-smb.fr

Réalisation : C. Brun, C. Roche
Collection « Terminologica »
ISBN : 978-2-919732-80-7
ISSN : 2607-5008
Dépôt légal : juillet 2020

Terminologie & Ontologie : Théories et Applications

Actes de la conférence

TOTh 2019

Le Bourget du Lac – 6 & 7 juin 2019

http://toth.condillac.org

avec le soutien de :
Université Savoie Mont Blanc
École d'ingénieurs Polytech Annecy Chambéry

Presses Universitaires Savoie Mont Blanc
Collection « Terminologica »

Comité scientifique

Président du Comité scientifique : Christophe Roche

Comité de pilotage

Rute Costa	Universidade Nova de Lisboa
Humbley John	Université Paris 7
Kockaert Hendrik	University of Leuven
Christophe Roche	Université Savoie Mont Blanc

Comité de programme 2019

Le comité de programme est constitué chaque année à partir du comité scientifique de TOTh en fonction des soumissions reçues. La composition du comité scientifique est accessible à l'adresse suivante : http://toth.condillac.org/committees

Guadelupe Aguado	Universidad Politécnica de Madrid – Spain
Amparo Alcina	Universitat Jaume I – Spain
Bruno Bachimont	Université Technologie de Compiègne – France
Jean-Paul Barthès	Université Technologie de Compiègne – France
Christopher Brewster	TNO – The Netherlands
Danielle Candel	CNRS, Université Paris Diderot – France
Sylviane Cardey	Université de Franche-Comté – France
Stéphane Chaudiron	Université de Lille 3 – France
Manuel Célio Conceição	Universidade do Algarve – Portugal
Rute Costa	Universidade NOVA de Lisboa – Portugal
Bruno Courbon	Université Laval – Canada
Lyne Da Sylva	Université de Montréal – Canada
Luc Damas	Université Savoie Mont-Blanc – France
Éric De La Clergery	INRIA – France
Dardo De Vecchi	Kedge Business School – France
Valérie Delavigne	Université Paris 3 – France
Sylvie Després	Université Paris 13 – France
Juan Carlos Diaz Vasquez	EAFIT University – Colombia
Hanne Erdman Thomsen	Copenhagen Business School – Denmark
Pamela Faber	Universidad de Granada – Spain
Christiane Fellbaum	Princeton University – USA
Iolanda Galanes	Universidade de Vigo – Spain
Christian Galinski	INFOTERM – Austria
François Gaudin	Université de Rouen – France
Teodora Ghiviriga	Alexandru Ioan Cuza University – Romania
Jean-Yves Gresser	ancien Directeur à la Banque de France – France
Ollivier Haemmerlé	Université de Toulouse – France
Gernot Hebenstreit	University of Graz – Austria
Amanda Hicks	University of Florida – USA
John Humbley	Université Paris 7 – France

Avant-propos

La Terminologie est une discipline scientifique à part entière qui puise à de nombreux domaines dont la linguistique, la théorie de la connaissance et la logique. Pour que cette diversité soit une richesse, il faut lui offrir un cadre approprié au sein duquel elle puisse s'exprimer et s'épanouir : c'est une des raisons d'être des Conférences TOTh créées en 2007. A ces conférences « mères » qui se tiennent chaque année à l'Université Savoie Mont-Blanc sont associées depuis 2011 les Journées d'étude TOTh dédiées à un thème plus spécifique organisées par une institution partenaire.

Dans ce contexte, la formation et la transmission des connaissances jouent un rôle essentiel. La *Formation TOTh* précédant la Conférence se déroule sur deux années consécutives dédiées pour l'une à la dimension linguistique et pour l'autre à la dimension conceptuelle de la terminologie, deux dimensions étroitement liées.

A la présentation de travaux sélectionnés par un Comité de programme international, la *Conférence TOTh* inclut une *Conférence invitée* et, selon les années, une *Disputatio*. La première, donnée par une personnalité reconnue dans son domaine vise l'ouverture à d'autres approches de la langue et de la connaissance. La seconde, à travers une lecture commentée effectuée par un membre du comité scientifique, renoue avec une forme d'enseignement et de recherche héritée de la scolastique.

Christian Galinski de Infoterm, a ouvert la conférence sur le sujet de « *The emergence of terminology science and terminological activities* ».

Cette année, comme en 2018, nous n'avons pas inclus de *Disputatio* par manque de temps. En effet, pour la première fois, TOTh a accueilli une session satellite, en parallèle avec la conférence, sur le thème de « Terminology and Text Mining » en lien direct avec les thèmes de TOTh. Nous avons également dédié une session de la conférence au projet Européen ELEXIS.

Les 29 communications et les 3 posters ont permis d'aborder de nombreux sujets tant théoriques que pratiques, autant d'exemples de la diversité et de la richesse de notre discipline. Je vous invite à découvrir à travers ces actes les 24 interventions qui ont donné lieu à publication.

Avant de vous souhaiter bonne lecture, j'aimerais terminer en remerciant tous les participants pour la richesse des débats et des moments partagés.

Christophe Roche
Président du comité scientifique

Sommaire

Conférence d'ouverture

The emergence of terminology science and terminological activities

Christian Galinski

1060 Wien, Esterhazygasse 12/1/21, Austria
christian.galinski@chello.at

Abstract. Today's terminology science has several roots and shows many approaches. In the course of development of modern sciences and technologies from the late 18th to the early 20th century scientists in many countries were active in designing nomenclatures in various domains. International cooperation and communication necessitated these to be unified and harmonized. The need for multilingual terminologies emerged in the course of early standardization efforts and finally when the official standards organizations were established. No wonder that engineers around and after 1900 engaged in large-scale terminological activities. They chose a practical-pragmatic concept-oriented approach in cooperation with many, sometimes thousands of experts, in order to develop multilingual terminologies. A comprehensive theory for these endeavours was developed only much later, after World War II. Besides, terminological activities diversified, and new approaches emerged. This contribution provides an overview on the development of the field of terminology with its broad range of practical activities and theoretical approaches.

1. Introduction

There are many theoretical and practical approaches dealing with "terminology". They developed out of many terminological activities that had different objectives and purposes evolving over time. "Terminology" emerges when it is necessary to communicate specialized knowledge and skills.

'Prescriptive approaches' particularly prevail in the natural sciences and technologies – which refers to the definition of concepts and assigning designations to them. It can also refer to the activities involved in this process, to the types of information (viz. metadata) to be used for creating terminological

entries, to the way how to use such entries in certain applications and other aspects.

'Descriptive approaches' prevail particularly in the social sciences and humanities where the use of terms is in the focus of investigation – for instance in special communication or in scientific-technical texts. Socioterminology for instance is concerned with social aspects of terminology evolution and use etc.

As ICT today have an impact on all social, economic, scientific and technical activities, the Council of German Language Terminology (RaDT:2017) positions "terminology" as follows showing the pervasive nature of terminology:

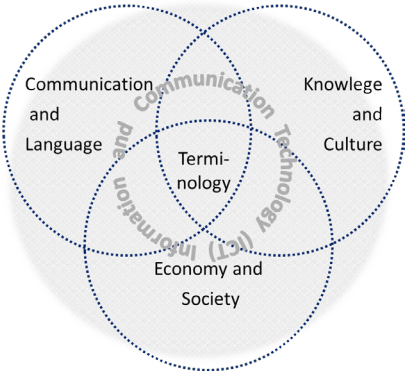


Fig. 1 Position of terminology in relation to other domains and subjects

The RaDT describes terminology science as the discipline focusing on concepts and their designations in specialized communication with the objective to describe terminological specificities (structures and processes) of specialized communication and knowledge transfer and to provide solutions to related problems.

The international standard ISO 704[1], clause 3.3 defines 'specialized concept' as *"concept which reflects specific or technical knowledge within a given subject field"*. More generally, the ISO 1087[2] defines 'concept' as a *"unit of knowledge created by a unique combination of characteristics"*. In the note 1 to entry it says *"Concepts are not necessarily bound to particular*

1 ISO 704:2009 "Terminology work – Principles and methods"
2 ISO 1087:2019 "Terminology work and terminology science – Vocabulary"

languages. They are, however, influenced by the social or cultural background, which often leads to different categorizations." As analyzing and categorizing thinking processes must have preceded resulting in the *"unique combination of characteristics",* concepts can also be taken as a unit of thought/thinking. And as knowledge also needs to materialize by means of representations for the purpose of interhuman communication, concepts can also be regarded as units of communication.

'Designation' in ISO 10241-1[3] is defined as *"representation of a concept by a sign which denotes it"* followed by two notes:

> *1 In terminology work three types of designation are distinguished: terms, symbols and appellations.*
> *2 Designations can be verbal or non-verbal or a combination thereof.*

In scientific-technical reality non-verbal representations at times may even replace a verbal definition or other kind of concept description. Besides, any language modality (accompanied or not by paralinguistic features or other non-verbal communication) can represent scientific-technical concepts.

The above coincides with Kalverkämper's (2013:57) scheme of "Taxonomics <culturemes and communication>" according to which communication can take place verbally in spoken, written or techno-medial form, and as non-verbal communication in vocal-paralinguistic or non-vocal body-language and non-vocal-pictorial form. ICTs increasingly can master all these modalities and their varieties – in terms of content input as well as different modality outputs – while the respective user communities (and their members) are increasing due to new needs.

Starting off with a historical background, this contribution focuses on the development of terminology science and terminological activities especially after 1850. It is based on the experience of the International Information Centre for Terminology (Infoterm), its relation to the technical committee ISO/TC 37 "Language and terminology" (ISO/TC 37:2004) and its involvement in a broad range of activities.

3 ISO 10241-1:2011 "Terminological entries in standards – Part 1: General requirements and examples of presentation"

2. Primordial terminology

From the perspective of special language and special communication research, terminologies emerged and exist already since prehistoric ages, latest since the Neolithic. (Knobloch 1998) They developed out of specialized practice, such as agriculture, military, building, shipbuilding etc. – of course also in early manufacturing developments for instance in the Roman and Chinese empires. This was inevitably accompanied by the development of technical terminologies and specialized languages that spread through cultural contact, peaceful trade or warlike conquests.

2.1. Primordial beginnings of terminology science

Human thinkers turned their interest on "terminology" – philosophically speaking in the meaning of representations of (scientific-technical) concepts – as early as during the development of ancient cultures about 2500 years ago. These early roots of philosophical thinking about concepts and terms continued to develop and were passed on to next generations slowly over many centuries in several languages and across cultural boundaries. Aristotle (384~322 BCE) for instance was keen to clarify concepts in all subjects he dealt with (Wenskus:1998) as was the Roman educator Quintilian (about 35~100). (Laurén:1998, p. 5ff.) The Romans preserved, imitated and spread these ideas over Europe until they were able to competitively rival the Greek culture. Latin language became widespread and the classical world became bilingual, Greek and Latin.

The Greco-Roman cultural foundation has been immensely influential on the languages, politics, law, educational systems, philosophy, science, warfare, poetry, historiography, ethics, rhetoric, art and architecture of the modern world. Grammatical terminology for example was shaped by the Greek, imparted by the Romans and largely exists in modern languages till today. (Funke 1999:2256) In this connection, it is noteworthy to recognize the crucial role of Arab philosophers, scientists and translators from the 9th to the 13th centuries in transmitting Greek, Hindu and other pre-Islamic knowledge to the Christian West.

During the "Age of Enlightenment" after the Middle Ages great advances took place in several field of philosophy, natural science and technology: They were mostly accompanied by terminological clarifications. Printing technology (invented around 1450) started to support national and international exchange of ideas which among others led to large-scale national specialized

encyclopaedic endeavours in several countries. A person usually referred to here is Denis Diderot (1713~1784), one of the most important organisers and authors of the French *Encyclopédie* which was a huge endeavour in all respects. Pioneering scientists to be mentioned here are Leonardo da Vinci (1452~1519), Nicolaus Copernicus (1473~1543), Galileo Galilei (1564~1642), René Descartes (1596~1650), Blaise Pascal (1623~1662), Gottfried Wilhelm Leibniz (1646~1716) – not to forget Isaac Newton and Denis Diderot. Newton's (1642-1727) publication *Principia Mathematica* (1687) is often regarded as the first major enlightenment work. Some thinkers of this period were also great scientific discoverers or technical inventors. The period also saw the beginnings of scientific nomenclatures and taxonomies, such as the major works of the Swedish botanist Carl Linnaeus (1707~1778) *Systema Naturae* (1735) and *Species Plantarum* (1753).

This development over more than centuries prepared the ground for the First Industrial Revolution when widely unified and harmonized – often multilingual – nomenclatures became an urgent need.

2.2. First and Second Industrial Revolutions

The spinning jenny invented by James Hargreaves in England in 1764 was one of the innovations that started the First Industrial Revolution which – at least in England – lasted until about 1870. The biggest changes came in the industries where mechanization started to replace agriculture by the industry as the backbone of the societal economy. The important invention of the steam engine provided the new type of energy that helped speed up manufacturing and development of railroads thus accelerating the economy. (Pouspourika 2019[4])

Around 1870 new sources of energy : electricity, gas and oil triggered the Second Industrial Revolution which lasted nearly hundred years. They enabled the early development of technical communication (i.e. telegraph and telephone) and the internal combustion engine, alongside with the exponential increase of the demands for steel and chemical products as well as means to transport goods and people. Research became more centralized, and capital focused on an economic and industrial model based on new "large factories"

4 https ://ied.eu/project-updates/the-4-industrial-revolutions/ accessed 2020-01-18

and the organizational models of production envisioned by Frederick W. Taylor (1856~1915) and Henry Ford (1863~1947)[5].

2.3. Establishment of formal standardization bodies

Following the signing of the "Metre Convention", the International Bureau of Weights and Measures (BIPM) was created in Paris, in 1875, as an intergovernmental organization. Given the high importance of harmonized quantities and units for industry, the BIPM today has the mandate to provide the basis for a single, coherent system of measurements throughout the world, traceable to the International System of Units (SI). In the field of inorganic chemistry an international conference was convened in Geneva in 1892 by the national chemical societies, from which the first widely accepted proposals for standardization arose. In 1919, the newly formed International Union of Pure and Applied Chemistry (IUPAC) took over these tasks and continues activities on a broad basis to this day. Because of the importance of electricity as the new source of energy the harmonization of the respective terminology became urgent. The first International Electrical Congress took place in Paris, in 1881. Following discussions at the 1900 Paris International Electrical Congress, the International Electrotechnical Commission (IEC) held its inaugural meeting on 26 June 1906. IEC was the second international standardizing body – the first one being the International Telegraph Union (now International Telecommunication Union, ITU) created in 1865 to set international standards in order to connect national telegraph networks.

By the end of the 19th century, differences in standards between companies was making trade increasingly difficult and strained leading to complaints, such as for instance by an English iron and steel dealer: "*Architects and engineers generally specify such unnecessarily diverse types of sectional material or given work that anything like economical and continuous manufacture becomes impossible. In this country no two professional men are agreed upon the size and weight of a girder to employ for given work.*"[6] The Engineering Standards Committee was established in London in 1901 as the world's first national standardizing body. It subsequently extended its standardization work and became the British Engineering Standards Association in 1918, adopting the name British Standards Institution (BSI)

5 https://pdfs.semanticscholar.org/03de/2f229d61f63b8c6fd5affe7d116d07ca0f10.pdf accessed 2020-01-18

6 https://en.wikipedia.org/wiki/Standardization accessed 2020-01-16

in 1931. The national standards were adopted universally throughout the country and enabled the markets to act more rationally and efficiently, with an increased level of cooperation. After the First World War, similar national bodies were established in other countries, such as 1917 in Germany (i.e. today's Deutsches Institut für Normung, DIN), followed by its counterparts in America (today's American National Standards Institute, ANSI) and France (today's Association française de normalisation, Afnor), both in 1918.

The definition and specification of quantities and units by the BIPM, the systematic elaboration and maintenance of electrotechnical terminology (i.e. International Electrotechnical Vocabulary, IEV) by IEC and the identification and description of about 200 m[7] identifiable chemical substances on the basis of the IUPAC rules (e.g. the *Gold Book* and the *Blue Book*) are huge endeavors and have a great impact on other systems, such as the international patent system including the "International Patent Classification" (IPC). The quantity of terminology contained in these and other nomenclatures and taxonomies may well near a staggering 500 m entries. (figure revised, see Galinski and Reineke 2011) Most of the existing nomenclature organizations started off from establishing a fundamental theoretical vocabulary and setting rules for the naming and presentation of the entities they harmonize. This example was followed by the emerging system of standardizing bodies around 1900.

The third international standards organization set up in the 1920s was the International Federation of the National Standardizing Associations (ISA). It was suspended in 1942 during World War II and newly established in October 1946 as International Organization for Standardization (ISO).

3. Roots of terminology science

The pioneering terminological activities among others of the Toledo School of Translators in the 12th century, the role of the Jesuits (after 1582) and others in China, the Japanese interpreters and translators in Nagasaki (1600~1868), the Italian Accademia della Crusca (est. 1582), followed by the Académie Française (est. 1635), and then the Royal Spanish Academy (est. 1713) and its later off-springs in Latin America[8] shall not be undervalued. The same applies to other activities already mentioned above. But until about 1850 thinking and activities rather followed speculation than science in today's

7 By CAS https://en.wikipedia.org/wiki/Chemical_Abstracts_Service accessed 2020-01-18

8 https://en.wikipedia.org/wiki/Royal_Spanish_Academy accessed 2020-01-18

sense – done by "dilettante" in the positive sense of the word which by no means excludes great advances and inventions made by individuals.

This chapter is on the philosophical (or philosophy of science) approaches to terminology, linguistic approaches and increasing efforts by scientific-technical communities to "professionalize" terminology work up to the Second World War.

3.1. Philosophy of science

As already mentioned, philosophers – especially those of science theory – as experts of their domain are constantly trying to clarify their terminology. The same applies to the fundamentals of all sciences, as it is a major aspect of epistemology and logic. "*A short glance at the history of science tells us: wherever special terminologies developed in scientific disciplines such as physics, botany, zoology, etc., they are based on classical conceptual logic. ... Prescriptive or normative terminology work thus becomes applied logic.*" (Oeser 1992)

According to E. Wüster (1974a), the fundamental theoretical study of the terminology of individual domains or subjects results in "specialized theories of terminology", while the comparative study of the patterns and structures of the terminologies of special domains or subjects leads to a "General Theory of Terminology" (GTT) or "General principles of terminology". A.D. Hajutin called this most abstract kind of terminology work "metaterminology" covering the study of logico-philosophical and general linguistic foundations for the construction of a model of specialized terms and term systems. (see Oeser:1992)

"*Since terminology is concerned with scientific concepts, rather than the concepts expressed in general language, there is a close link to the philosophy of science, founded by the Vienna Circle. ... In particular Carnap's book on the logical structure of the world in 1928 already contained links to the theory of terminology based on the idea that a "logical framework" is necessary to "map" the world in language (see Nedobity 1984[9], p.45). This idea had also been developed by Wittgenstein in his Tractatus logico-philosophicus, which had a fundamental influence on the philosophy of the Vienna Circle.*" (Oeser 1992)

9 Nedobity, Wolfgang. Eugen Wüster und die Sprachkritiker des Wiener Kreises. In M*uttersprache* 95(1984/85)1/2, 42~48

"*Although the philosophy of science of the Vienna Circle and Wüster's theory of terminology had the same fundamental idea, there is an important difference. ... Wüster started from conceptual logic. For the Vienna Circle science was a propositional system, for Wüster a concept system. ... Physics was the model discipline, where Mach, Boltzmann and Hertz had already laid the basis for its metatheoretical foundation. ... This led to the programme of the analysis of the logical syntax of scientific language, ... although Gödel had shown by the example of the elementary theory of numbers that, even in mathematics, it is impossible to prove absolute consistency.*" (Oeser 1992)

The development of scientific languages had started without established rules and principles in the 18th and 19th centuries. "*The consequence was chaos in the concept systems of almost all scientific disciplines at the dawn of the modern era. In physics there was chaos in measurement systems, in biology, chaos not only in designating special concepts, but also concerning distinctive characteristics used for the classification of plants and animals. ... In all subject fields, the natural systems of concepts start to proliferate with an unordered multitude of terms which can be reduced only later.*" (Oeser 1992)

"...Carnap's formalistic philosophy of science sought liberation from factual content through a logical reconstruction of scientific language, transforming factual object-oriented propositions I not a formalism in order to check for inconsistency..." (Oeser 1992) The reaction of Wüster and Wittgenstein to basic positions of the Vienna Circle resembled insofar as both applied "language criticism". While Wittgenstein focussed on criticism of general language, Wüster dealt with criticism of special language. "*With conceptual logic being prior to propositional logic, Wüster rejected a widespread opinion, also supported by the Vienna Circle, that modern formalized logic has completely replaced classical logic. ... as a reference for the fundamental non-replaceable role of classical conceptual logic, [he refers to] the logician von Freytag-Löringhoff (1955), who proved, more consistently than other defendants of classical logic, that conceptual logic contains formalized propositional logic as a special case.*" (Oeser 1992)

Somehow, the conception of science as a concept system (Wüster) and as a propositional system (Carnap) complement each other. While the "linguistic turn" had influenced philosophy of science in the decades before and after 1900, the transition from early language-focused approaches to conceptual logic-oriented terminology research can also be called the philosophy of science turn of terminology science. (see Oeser 1992)

3.2. Linguistics and terminology theory

As already mentioned, philosophers – especially those of science theory – since earliest beginnings have also reflected about "language" and in this connection dealt with terminology. The "linguistic turn" in philosophy of science is said to have started with Gottlob Frege's work *The Foundations of Arithmetic* (1884) specifically paragraph 62 where he explores the identity of a numerical proposition. The concern for the logic of propositions and their relationship to "facts" was later taken up by the notable analytic philosopher Bertrand Russell in *On Denoting*. His associate Ludwig Wittgenstein was one of the progenitors of the linguistic turn.[10]

A second great impact on the contemporary philosophy of science came from the Swiss Ferdinand de Saussure (1857~1913) whose ideas laid a foundation for many significant developments in both linguistics and semiology in the 20th century. He is widely considered one of the founders of 20th-century linguistics and one of two major founders of semiotics/semiology together with the American Charles S. Peirce (1839~1914).[11] The latter was educated as a chemist and became a renowned philosopher, logician, mathematician – sometimes known as "the father of pragmatism". Both had an influence on applied linguistics – especially on translation studies, here in particular specialized translation where terminology plays a crucial role.

A third impact in the first quarter of the 20th century is due to the increasing internationalization of the sciences, industry and trade which necessitated the introduction of specialized language education and training especially at specialized universities and colleges particularly in Germany (and neighboring countries) and the Nordic countries. This triggered the need to study specialized languages (German: Fachsprachen) and specialized communication (German: Fachkommunikation).

One major representative of this direction is the "Prague School of Terminology" – or the Prague School of the specialized language of economics, short: Prague School of Linguistics. It considered specialized languages as functional language systems based on a scientific concept system, whereby the totality of terminological and non-terminological linguistic means constitutes this functional system which comprises terminological units, specialized syntactic phrases (German: Fachwendungen), non-terminological elements

10 https://en.wikipedia.org/wiki/Linguistic_turn accessed 2020-01-18

11 https://en.wikipedia.org/wiki/Ferdinand_de_Saussure and https://en.wikipedia.org/wiki/Charles_Sanders_Peirce accessed 2020-01-18

and non-specialized language phrases (German: nicht-fachsprachliche Wendungen). (Felber/Budin:1989, 46ff.) The Prague School strongly influenced special language and education especially in the Nordic countries and in the Eastern European countries after 1945.

A less visible fourth impact comes from experts and theoreticians of planned languages, especially Esperanto.

Wüster did research on and was a practical expert in nearly all of the above-mentioned approaches of philosophy of science as well as linguistic approaches – including also general and specialized lexicography. This can be gathered from his extensive correspondence with experts in many countries, the books and other materials collected in his private library and numerous articles. (Felber:1998) His activities in associations and committees is another indication for his efforts to build theoretical-methodological bridges between the above-mentioned approaches – later including also classification studies and I&D (information and documentation).

3.3. Terminological activities by expert communities until 1938

The strongest impulses for systematic terminological activities emanated from the emerging scientific and technical domains during the First and Second Industrial Revolution in the wake of other developments, such as:

- Latin as lingua franca lost ground making way for national languages
- Mechanization of production transformed into industrial production
- Need for standardization and harmonization increased
- Specialized knowledge – and thus also new concepts and terms – started to grow exponentially and continues to do so today. (see Laurén:1998, p.6ff.)

Part of these impulses were absorbed by the upcoming standardizing activities, which from the very beginning also required the standardization of fundamental terminology. Part of them grew into big international organizations (such as IUPAC) – or were taken as a model for highly authoritative organizations (such as WIPO[12]).

Whenever a systematic approach is chosen for terminology work, it needs to be based on explicit principles and methods. No wonder that the

12 https://en.wikipedia.org/wiki/World_Intellectual_Property_Organization accessed 2020-01-18

prescriptive approach in terminology science emerged in the field of technical standardization largely involving technical domain experts.

Soon after 1900 the Association of German Engineers contracted an expert of German language studies for the task of collecting all terminology existing in the German language in the "VDI-Technolexikon". 1907 the first results were evaluated and found that it would need another 40 years to complete this compilation, if the alphabetical approach chosen were continued. Discovering that Schlomann's new method of classified order is superior to traditional lexicographical methods, the VDI stopped work on the Technolexikon and started to support Schlomann. 1907~1932 Schlomann published – in international cooperation with trained 'terminologists' – 17 *Illustrierte Technische Wörterbücher* in six languages in different subject fields based on guidelines conceived by Schlomann. (Felber/Budin :1998, 140) They followed a classified order and contained many figures.

IEC started 'systematic' work on terminology in 1908 but switched to a structured approach in 1927 (1st edition 1938). The efforts on the IEV are accessible today online through IEC's "Electropedia".[13] In 1936, triggered by the publication of E. Wüster's (1931) dissertation *Internationale Sprachnormung in der Technik* [International standardization of technical language], ISA established a Technical Committee ISA/TC 37 "Terminology" to formulate general principles and rules for terminology standardization. ISA planned four classes of future recommendations :

1. Vocabulary of terminology
2. Procedure for preparing national or international standardized vocabularies
3. National and international standardization of concepts, terms and their definitions : principles for their establishment and criteria of value
4. Layout of monolingual and multilingual vocabularies, including lexicographical symbols.

4. Terminological activities after 1945

ISO was established in 1946 and officially began operations in February 1947 – one of the technical committees foreseen from the very beginning was ISO/TC 37 "Terminology (principles and co-ordination)". In 1951, Wüster

13 https ://en.wikipedia.org/wiki/International_Electrotechnical_Vocabulary accessed 2020-01-18

(with great efforts involving international organizations) saved ISO/TC 37 from being disbanded due to inactivity. The committee became operational in 1952 and continued with an adapted ISA scheme of planned recommendations (ISO/R). It took more than 15 years until five ISO/Rs and one ISO standard were published. (ISO/TC 37:2004) During this time, Wüster managed the committee at his private terminology centre in Wieselburg (Lower Austria) on behalf of the Austrian Standards Institute (today's ASI).

Wüster continued his great efforts to promote the field of terminology and to realize an old dream: the establishment of an international terminology centre. In 1969, UNESCO launched the UNISIST Programme, aiming at the exchange of scientific information at the international level. (see Report 1971[14]) Wüster was asked to submit the following two reports:

1. Inventory of sources of scientific and technical terminology
2. A plan for establishing an international information centre (clearing house) for terminology. (published in Wüster:1974b)

In addition, the Council of Europe showed interest in the coordination of terminology training and contacted Eugen Wüster on that matter. These more or less concerted efforts finally led to the establishment of Infoterm in 1971 – at a time, when ISO/TC 37 started its consolidation phase. H. Felber became the director of Infoterm and secretary of ISO/TC 37.

This development coincides with the beginning of the Third Industrial Revolution. With the emergence of a new type of energy appeared whose potential surpassed its predecessors: nuclear energy. This revolution witnessed the rise of electronics (incl. the transistor and microprocessor) and the rise of telecommunications and computers. New technology also emerged from space research and biotechnology and gave rise to the era of high-level automation in production.

Being relieved from some of his organizational obligations, Wüster concentrated on working on several sub-topics of his approach. In 1972 he was asked to teach a course on "Lexicology and lexicography with a special focus on terminology theory and language standardization" at the Institute of Linguistics of the University of Vienna. He took the just published ISO/Rs as a basis but further developed this material in the direction of his main interest. The unfinished manuscript of this course was published posthumously after careful editing by H. Felber. (Wüster:1979)

14 https://unesdoc.unesco.org/ark:/48223/pf0000064862 accessed 2020-01-18

In the course of the 1970s – not least due to the influence of UNESCO – experts of I&D entered standardization activities in ISO/TC 37. Given the fact that terminology work is a peculiar I&D activity, namely recording factual data on concepts and their representations, Infoterm continued promoting the interoperability of terminology theory and work with the field of I&D especially in the 1980s and 1990s. (Galinski 2019) This lead among others to the conception of Terminology and Documentation (T&D) as a prerequisite of knowledge management.

During the 1980s there were big discussions about the Vienna, Nordic, Soviet, Prague, German, Canadian and other "schools of terminology" on the one hand and disputes about the onomasiological vs. semasiological approach on the other hand. In the course of the years, the discussions about different schools and national/regional approaches yielded to the recognition that different socio-economic, political or societal conditions require different adaptations of terminological principles and methods. This also applies to different organizational environments. In Europe, experts of different disciplines worked together with terminology experts in many EU-projects. This facilitated the acknowledgement of the *"plurality of theoretical approaches in terminology"*. (Costa :2006) In fact, such plurality is a phenomenon not unfamiliar to many other disciplines.

After 1980, language aspects, such as computational terminography, multilingualism, localization (L10N), language for specific purposes (LSP), terminology planning, specialized lexicography, language resource management (2002), translation and interpretation (2012), computer-assisted and automatic translation, ontology (building and engineering), etc. entered the sphere of Infoterm activities and found their way into standardization activities in ISO/TC 37.

Via EU-projects ISO/TC 37 standards had an impact on several other standardization activities in the fields of metrology, traffic informatics, Internet of Things (IoT), product classification and master data management, and lately also on activities with respect to eAccessibility and eInclusion. Everywhere, where structured content at the level of lexical semantics – i.e. small content entities data-modelled according to metadata (also called microcontent) – is used, the terminology approach which is multilingual and multimodal from the outset proves a good model to be followed. This applies to all eApplications, such as eBusiness, eHealth, eLearning, etc.

5. Outlook

After intensive discussions at the RaDT (2017[15]), the "old" semantic triangle as adopted by Wüster was extended:

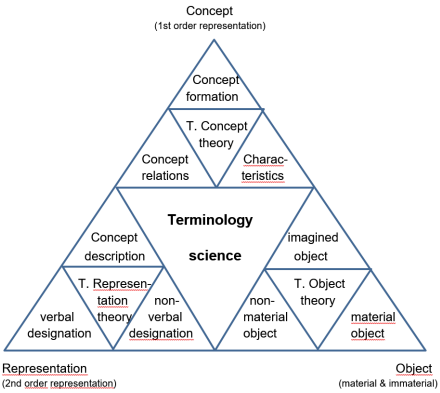


Fig. 2: extended semantic triangle

Analysing the extended triangle under the aspect of "completeness" of terminology science approaches one must state that

(1) The edge of "representation" is most developed, especially from a linguistic point of view, but non-linguistic (spoken, visual and other) language modalities and alternative communication means (incl. sign languages etc.) are still under-represented.

(2) The "concept edge" (1st order representation) is still much under-developed and definitely needs the integration of findings of recent brain research, possibly also of IoT's "virtual objects" that cover among others scientific-technical concepts. (ITU[16])

(3) The "object edge" lacks considering developments with respect to structured content in eBusiness (e.g. product master data), IoT, etc.

At each edge one can have different views if seen from the other two edges. Any data occurring in this triangle is an element of structured content (or structured data). Under the perspective of integration and interoperability it is

15 http://radt.org/images/veroeffentlichungen/Wissenschaft%20-RaDT_2016_rz_16seiten.pdf acc. 2020-01-18

16 ITU-T Y.2060:2012 https://www.itu.int/ITU-T/recommendations/rec.aspx?rec=y.2060 acc. 2020-01-18

imperative that they are governed by – more and more harmonized – metadata in order to acquire the (potential) capability of content interoperability across language and system boundaries. The situation today somehow resembles the problem of "interchangeability" of machine elements to be achieved without human intervention after their production with improved machine tools during the First Industrial Revolution. For the sake of comprehensive interoperability, today's ICT tools need substantial upgrading based on standards. The metadata for structured content need further development and harmonization based on standards.

References

Costa, Rute. 2006. "Plurality of theoretical approaches to terminology". In *Modern approaches to terminological theories and applications* edited by Heribert Picht, 77~89. Bern : Peter Lang.

Felber, Helmut ; Budin, Gerhard. 1989. *Terminologie in Theorie und Praxis* [Terminology in theory and practice]. Tübingen : Gunter Narr.

Felber, Helmut. 1998. "Eine erweiterte Wüster Bibliographie (1931~1977)" [An extended Wüster bibliography (1931~1977]. In *Eugen Wüster (1898~1977). His life and work. An Austrian pioneer of the information society* edited by Erhard Oeser and Christian Galinski, 235~323. Vienna : TermNet.

Funke, Hermann. 1999. "*Grammatik, Rhetorik und Dialektik (Trivium) und ihre Fachsprachen : eine Übersicht"* [Grammar, rhetorics and dialectics (Trivium) and their specialized languages : an overview]. In *Languages for Special Purposes. An International Handbook of Special Language and Terminology Research* edited by Lothar Hoffmann, Hartwig Kalverkämpfer, Herbert Ernst Wiegand, 2255~2260. Berlin/New York : de Gruyter.

GALINSKI, Christian ; REINEKE, Detlef. "Vor uns die Terminologieflut" [Facing the terminology deluge]. 8~12. *edition*, 7(2011)2. ISSN 1862-023X

Galinski, Christian. 2019. „Blütezeit der Zusammenarbeit zwischen Terminologie einerseits und Information und Documentation (IuD) andererseits : 1980~2000" [Heyday of cooperation between terminology and I&D : 1980~2000]. In *Terminologie : Epochen, Schwerpunkte, Umsetzungen. Zum 25-jährigen Bestehen des Rats für Deutschsprachige Terminologie* edited by Petra Drewer, Donatella Pulitano, 21~43. Berlin : Springer Nature.

ISO/TC 37. 2004. *50 Years ISO/TC 37 Terminology and other language resources – A history of 65 years of standardization of terminological principles and methods*. Geneva : ISO (ISO/TC 37 N499)

Kalverkämper, Hartwig. 2013. "Körperkommunikation als Teil von Translationskultur" [Body communication as part of translation culture]. In Arbeiten zur Theorie und Praxis des Übersetzens und Dolmetschens 63 edited by Klaus-Dieter Baumann, Hartwig Kalverkämper, Klaus Schubert, 51~113. Berlin : TransÜD.

Knobloch, Johan. 1998. „Fachsprachenforschung in vorhistorischen Sprachen : Forschungsansätze und Sprachrelikte" [Special language research on prehistoric languages : Research approaches and language relics]. In *Languages for Special Purposes. An International Handbook of Special Language and Terminology Research* edited by Lothar Hoffmann, Hartwig Kalverkämpfer, Herbert Ernst Wiegand, 289~295. Berlin/New York : de Gruyter.

Laurén, Christer ; Myking, Jan ; Picht, Heribert. 1998. *Terminologie unter der Lupe. Vom Grenzgebiet zum Wissenschaftszweig [Terminology under scrutiny. Development from an interdisciplinary subject to a discipline of science]*. Vienna : TermNet.

Oeser, Erhard. 1994. "Terminology and philosophy of science". In *International Conference on terminology science and terminology planning Riga, 17~19 August 1992. International IITF Workshop Theoretical issues of terminology science. Riga, 19~21 August 1992* edited by Jennifer K. Draskau, Heribert Picht, 24~34.Vienna : TermNet.

Paxton, John. "Mr. Taylor, Mr. Ford, and the Advent of High-Volume Mass Production : 1900-1912". *Economics & Business Journal : Inquiries & Perspectives*. 4(2012)1, 74~90

Wenskus, Otto. 1998. "Reflexionen zu fachsprachlichen Phänomenen in der Antike und Spätantike" [Reflections on special language phenomena in antiquity and late antiquity]. In *Languages for Special Purposes. An International Handbook of Special Language and Terminology Research* edited by Lothar Hoffmann, Hartwig Kalverkämpfer, Herbert Ernst Wiegand, 295~301. Berlin/New York : de Gruyter.

Wüster, Eugen. 1936. *Internationale Sprachnormung in der Technik. Besonders in der Elektrotechnik* [International standardization of technical language. Especially in electrotechnology]. Berlin : VDI, 1931

Wüster, Eugen. 1974a. „Die Allgemeine Terminologielehre – Ein Grenzgebiet zwischen Sprachwissenschaft, Logik, Ontologie, Informatik und den Sachwissenschaften“ [General Theory of Terminology – A border field between linguistics, logic, ontology, information science and the subject fields]. 61~106. *Linguistics* (1974)119.

Wüster, Eugen. 1974b. *The road to Infoterm*. München : Verlag Dokumentation (Infoterm Series 1)

Wüster, Eugen. 1979. *Einführung in die Allgemeine Terminologielehre und Terminologische Lexikographie* [Introduction to the General Theory of Terminology and terminological lexicography]. Vol. 1&2. Wien/New York : Springer-Verlag (TU Vienna Series Vol. 8)

Résumé

Les fondements de la Terminologie actuelle en tant que science sont divers tout comme les approches auxquelles elle donne lieu. Au cours du développement des sciences et des technologies modernes de la fin du XVIII^{e} siècle au début du XX^{e} siècle, les scientifiques de nombreux pays ont participé activement à la conception de nomenclatures dans divers domaines. La coopération et la communication internationales ont nécessité l'unification et l'harmonisation de ces nomenclatures. Le besoin de terminologies multilingues est apparu au cours des premiers efforts de normalisation et finalement lors de la création des organismes officiels de normalisation. Il n'est donc pas étonnant que les ingénieurs, à partir de 1900, se soient engagés dans des activités terminologiques à grande échelle. Ils ont opté pour une approche pratique et pragmatique, axée sur les concepts, en coopération avec de nombreux, parfois des milliers d'experts, afin de développer des terminologies multilingues. Une théorie complète de ces efforts n'a été élaborée que beaucoup plus tard, après la Seconde Guerre mondiale. Par ailleurs, les activités terminologiques se sont diversifiées et de nouvelles approches ont émergé. Cette contribution donne un aperçu du développement du domaine de la terminologie avec son large éventail d'activités pratiques et d'approches théoriques.

Articles

Étude comparative de deux méthodes outillées pour la construction de terminologies et d'ontologies

Sylvie Després*, Christophe Roche** ***, Maria Papadopoulou** ****

Université Paris 13,
Sorbonne Université, Inserm, Laboratoire LIMICS, Bobigny
(France) - http://www.limics.fr/
** Équipe Condillac « Terminologie & Ontologie » - Laboratoire LISTIC
Université Savoie Mont-Blanc (France) - www.condillac.org
*** Knowledge Engineering & Terminology Research Centre
University of Liaocheng (China) - www.ketrc.com

Résumé. Cet article propose de comparer deux approches outillées pour la construction de terminologies dont le système conceptuel est une ontologie formelle. Le premier, Protégé, est le logiciel le plus utilisé pour la construction et la maintenance d'ontologies. Logiciel libre, il est supporté par une très large communauté d'utilisateurs. Sa théorie du concept (classe) repose sur la logique des descriptions. Le second, Tedi, est une méthode outillée de construction d'ontoterminologies qui se veut en accord avec la façon de penser des experts. Il repose sur une théorie du concept compatible avec les normes ISO en Terminologie. Le domaine d'application choisi relève des Humanités Numériques. Notre exemple se limite à la définition de quelques termes désignant différents types de vases de la Grèce antique. Les critères de comparaison pris en compte portent sur le logiciel lui-même (architecture, disponibilité, ergonomie, interopérabilité, documentation), sur la dimension conceptuelle (théorie du concept), la dimension linguistique (théorie du terme), la méthodologie (construction de la terminologie et de l'ontologie) et enfin le point de vue de l'expert (prise en main, autonomie, réponses à ses besoins).

1. Introduction

Dans la mesure où, dans le cadre de cet article, nous nous intéressons principalement à définir les termes au regard des choses (concepts) qu'ils

désignent (définitions terminologiques), et non à leur signification en discours (définitions lexicographiques), nous considérerons la Terminologie sous l'angle de l'ISO telle que l'ISO la définit à travers les normes transverses ISO 1087-1 (ISO 1087-1) et ISO 704 (ISO 704).

En posant que le terme est une « désignation verbale d'un concept général » (ISO 1087-1), nous insistons non seulement sur la double dimension, linguistique et conceptuelle, de la Terminologie, mais aussi sur une certaine primauté du concept. Définir, c'est alors représenter un concept par « un énoncé descriptif permettant de le différencier des concepts associés » (ISO 1087-1). Le problème devient alors celui de la représentation du concept.

Aujourd'hui, la notion d'ontologie, au sens de l'Ingénierie des Connaissances, constitue une des perspectives les plus prometteuses pour la représentation du système conceptuel (Roche, 2005). L'ontologisation de terminologies (Papadopoulou & Roche 2018) ouvre la voie à de nombreuses applications : moteurs de recherche sémantique multilingues, dictionnaires électroniques, gestion d'information, etc. Il en est de même de la construction de termino-ontologies où un modèle conceptuel du domaine d'application utilisant des ressources terminologiques est élaboré afin d'être formalisé et opérationnalisé (Desprès, 2016, 2018). L'objectif est la construction d'artefact répondant à des besoins utilisateurs (Azzi *et al*., 2018), (Dandan *et al*., 2018).

L'ontologie[1], en explicitant le concept, pose non seulement le problème de sa construction et de sa mise en relation avec la dimension linguistique, mais aussi celui de la « mise en correspondance » de la théorie du concept sur laquelle elle repose avec la théorie du concept en Terminologie, pour autant qu'on arrive à s'accorder sur cette dernière. Rappelons que la Terminologie, au sens où nous la considérons ici, définit un concept comme une « unité de connaissance créée par une combinaison unique de caractères » (ISO 1087-1), structure qui, insistons sur ce point, se veut proche du mode de pensée des experts du domaine, et qu'on retrouvera dans la définition aristotélicienne du terme.

Cet article propose de comparer deux approches outillées pour la construction de terminologies dont le système conceptuel est une ontologie formelle. Le domaine d'application choisi relève des Humanités Numériques. On s'in-

1 définie comme une spécification dans un langage compréhensible par un ordinateur d'une conceptualisation, c'est-à-dire d'un système de concepts liés par des relations (Gruber, 1992), (Guarino *et al*., 2009) pour n'en citer que deux.

téresse ici à l'expert devant définir des termes désignant différents types de vases de la Grèce antique.

Les critères de comparaison pris en compte portent sur le logiciel lui-même (architecture, disponibilité, ergonomie, interopérabilité, documentation), sur la dimension conceptuelle (théorie du concept), la dimension linguistique (théorie du terme), la méthodologie (construction de la terminologie et de l'ontologie) et enfin le point de vue de l'expert (prise en main, autonomie, réponses à ses besoins).

L'article est organisé de la façon suivante. Le chapitre deux présente le domaine d'application, les besoins terminologiques de l'expert et les ressources disponibles. Les deux chapitres suivants sont dédiés à la présentation des deux environnements. Le cinquième chapitre dressera un comparatif des deux outils qui sera résumé dans la conclusion.

2. Les vases de la Grèce antique

La céramique constitue une des expressions les plus concrètes et les mieux connues de la civilisation grecque. De nombreux vases furent retrouvés dès le XVIII^{e} siècle provenant aussi bien de maisons particulières que de sanctuaires ou nécropoles (vases déposés en offrandes). La céramique contribua, à travers la diffusion d'images clairement identifiables sur les poteries, à une meilleure connaissance de la civilisation grecque dans ses dimensions aussi bien religieuse et mythologique (panthéon grec, mythes, épisodes des poèmes homériques, etc.) que de la vie quotidienne (vêtement, mobilier, monde féminin, vie dans la cité, banquet, etc.), ou la guerre (armement, combat), l'artisanat, les pratiques cultuelles, le sport, l'amour, etc.

Figure 1 : Cratère en calice représentant Poséidon et Thésée.

Nous nous sommes intéressés à la céramique athénienne dont la suprématie s'est exercée dès 550 av. J.-C. dans le monde grec, le quartier du Céramique à Athènes devint alors l'unique centre de fabrication, Athènes exportant largement sa production. Parmi les différents types de poteries et les différentes périodes qui se sont succédé, nous avons choisi de nous intéresser plus particulièrement aux vases de l'époque archaïque et classique, en s'appuyant sur différentes sources provenant, entre autres, des archives Beazley du Centre de Recherche sur l'Art Classique de l'Université de Oxford (Beazley 2019), des livres « Athenian Black Figure Vases » (1974) et « Athenian Red Figure Vases » (1975) de John Boardman, le livre « Athenian Vase Construction. A Potter's Analysis » de Toby Schreiber (1999), le livre « Greek Painted Pottery » de Cook R. M. (1997), le « Visual Glossary of Greek Pottery » de Cartwright, M. (2013), le livre de G. Richter et Marjorie Milne (1935) « Shapes and Names of Athenian Vases », ainsi que du Thesaurus AAT Getty (AAT Getty 2019).

Dans le cadre de cette étude comparative nous nous sommes limités à la définition des concepts correspondant aux types de vases désignés par les termes « cratère », « cratère à colonnettes », « cratère à volutes », « cratère en cloche » et « cratère en calice ». Rappelons que nous sommes dans un contexte où les termes sont connus des experts, il n'y a donc pas d'étapes préalables d'extraction et de sélection de candidats termes.

Nos hypothèses de travail sont les suivantes :

1) Définition des termes : *Définition de chose*, la définition des termes est une définition aristotélicienne en genre prochain et différences spécifiques correspondant aux caractéristiques essentielles du concept dénoté par le terme.
2) Définition des concepts : Un concept est défini comme une combinaison unique de caractéristiques essentielles, suivant en cela les principes des normes ISO. Par caractéristique essentielle on entend une caractéristique telle que, retirée de la chose, la chose n'est plus ce qu'elle est. Ainsi, *with column-like handles* est une caractéristique essentielle du concept désigné par le terme « column krater » (« cratère à colonnettes »).

Pour la définition des concepts correspondants aux types de cratères que nous avons cités, nous nous sommes appuyés sur leur description et sur la définition des termes tirée des archives Beazley que nous reproduisons ci-dessous :

> The term 'krater' suggests a mixing-vessel (compare Greek kerannumi - to mix), and we know that the wine served at the symposium was mixed with water. On vases decorated with symposium-scenes, a large open container with a foot is often depicted, and the name krater is appropriate. Examples can be traced back to the large Geometric examples that were used as grave-markers, and this funerary connection continues to be important. Excavations of burial-sites have shown that they could be used in Greek settlements overseas as containers of ashes, and South Italian, especially Apulian, volute-kraters often carry explicit funerary iconography. In the Athenian repertoire, there are four main types identified today : column-, volute-, calyx- and bell-. The psykter is a short-lived shape, used to cool the liquid. It is discussed here because it is often shown being used in kraters.

Column-krater : Named for its column-like handles, the column-krater is first known from Corinthian examples dated to the late seventh century. It is regularly produced by Athenian potters from the first half of the sixth-century until the third quarter of the fifth. It seems from graffiti on Athenian red-figure examples that the vessel was referred to as *Korinthios* or *Korinthiourges*.

Volute-krater : The volute-krater is named after its handles. The François Vase is a famous and early example, but the typical Athenian form occurs only later in the sixth century, with the handles tightly curled so that they look like the volutes on Ionic columns. The shape is also found in metal. Over the course of the fifth and fourth centuries, examples become slimmer, and Apulian volute-kraters from South Italy are particularly elaborate.

Calyx-krater : The handles of the calyx-krater are placed low down on the body, at what is termed the cul. Their upward curling form lends the shape an appearance reminiscent of the calyx of a flower, hence the name. The earliest known example was possibly made by Exekias in the third quarter of the sixth century. It continues to be produced, mainly in red-figure, becoming more elongated over the course of the fifth and fourth centuries.

Bell-krater : The latest of the four krater-types, it first occurs in the early fifth century, and is not found decorated in black-figure. It is named for its bell-like shape, perhaps originating in wood. It has small horizontal upturned handles just over halfway up the body. Some do not have a foot, and earlier examples may have lugs for handles. Over the course of the fifth and fourth centuries, the shape becomes slimmer.

3. Protégé

Dans la première approche la définition de l'ontologie adoptée est celle proposée par (Studer *et al*., 1998) : « une spécification formelle explicite d'une conceptualisation partagée ». Cette approche est fondée sur les activités intervenant dans le cycle classique de construction d'ontologies (scénario 1 de la méthodologie Neon) commun à toutes méthodologies de construction [Suárez-Figueroa *et al*., 2015]. Il est constitué des étapes de spécification, planification, conceptualisation, formalisation, implémentation. Au cours de ce cycle, les activités mises en œuvre (Suárez-Figueroa, M.C., Gomez-Perez, A., 2008) comportent les activités de gestion, orientées développement (pré-développement, développement, post-développement) et de support à la construction. Dans la phase de conceptualisation les principes différentiels pour la construction de la hiérarchie des concepts sont appliqués (Bachimont, 2000).

Protégé (Musen, 2015) est l'éditeur sélectionné pour construire la ressource termino-ontologique. Il est développé en Java, gratuit et son code source est publié sous une licence libre (la Mozilla Public License). Nous utilisons Protégé 5 qui permet la construction et la gestion d'ontologies OWL2. L'utilisation de l'éditeur et des plugins associés permet de réaliser la plupart des étapes de construction d'une ontologie.

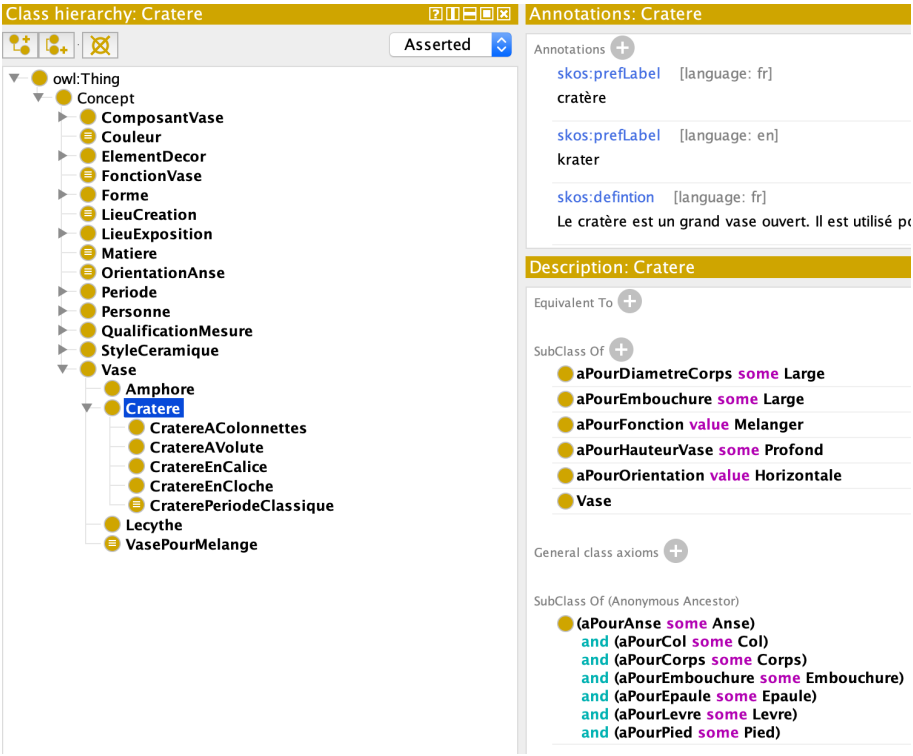


Figure 2. Éditeur Protégé.

4. Tedi

Si l'ontologie constitue aujourd'hui une des perspectives les plus intéressantes pour la Terminologie dite « conceptuelle », en particulier à des fins de traitement de l'information, la construction de l'ontologie avec des outils tels que Protégé demeure problématique. De l'aveu même de ses concepteurs, l'utilisation d'un tel environnement reste difficile sans l'aide d'un ingénieur de la connaissance : "As the group that developed Protégé, the most widely used ontology editor, we are keenly aware of how difficult the users perceive this task to be." (Horridge *et al.*, 2013). Le deuxième environnement que nous présentons, Tedi, vise à réduire le "fossé" qui peut exister entre les experts du domaine et ce type d'outils (Roche and Papadopoulou 2019).

Tedi[2] (ontoTerminology Editor) est une méthode outillée de construction d'ontoterminologies multilingues, terminologies dont le système conceptuel est une ontologie formelle (Roche, 2007). Tedi met en œuvre la CTT (Concept Theory for Terminology (Roche 2020)), une théorie du concept dédiée à la Terminologie qui se veut proche de la façon de penser des experts (définition aristotélicienne en genre prochain et différence spécifique). La CTT repose sur les notions de *caractéristiques essentielles* et de *caractéristiques descriptives*, à partir desquelles se définissent les concepts : un concept étant défini comme une combinaison unique de caractéristiques, suivant en cela les recommandations des normes ISO en Terminologie.

Tedi met à la disposition des experts un ensemble d'éditeurs dédiés à la dimension conceptuelle (éditeurs de concepts (figure 3), d'objets, de caractéristiques essentielles et descriptives, de relations) et à la dimension linguistique (éditeurs de termes et de noms propres dans les différentes langues (figure 4)).

La méthodologie associée à Tedi combine les approches linguistique et conceptuelle et lui dédie un éditeur spécifique. Cette méthodologie repose sur l'idée centrale qu'un concept est un ensemble de caractéristiques essentielles suffisamment « stable » pour être nommé en langue par un terme (Roche et Papadopoulou 2019). L'identification préalable des termes permet de guider la définition des concepts, à condition d'avoir également au préalable identifié les caractéristiques essentielles constitutives de tout concept : *with-neck*, *without-neck*, *with-column-handles*, *with-upward-curling-handles*, etc. On assiste ainsi à un retour de l'importance du terme dans une démarche principalement conceptuelle : la langue participe au découpage du réel en concepts, ensembles « stables » de caractéristiques, même si la conceptualisation du domaine peut amener, principalement pour des raisons d'organisation de la connaissance, à la définition de concepts sans désignation en langue (terme), par exemple le concept <Vessel with neck> [Roche 2020]. Enfin, Tedi guide l'utilisateur dans la construction de l'ontologie en ne proposant que des choix possibles en accord avec la théorie du concept sous-jacente. Les vérifications sont faites « à la volée » et non pas *a posteriori*, ce qui constitue une aide importante pour l'expert.

2 http://ontoterminology.com/tedi

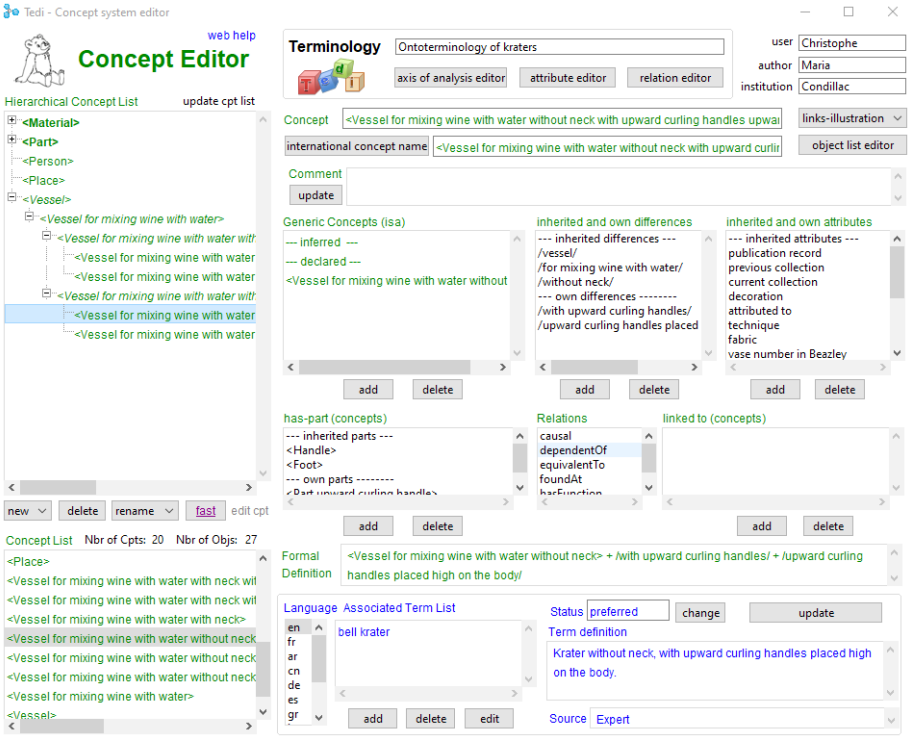


Figure 3. Éditeur de concepts de Tedi

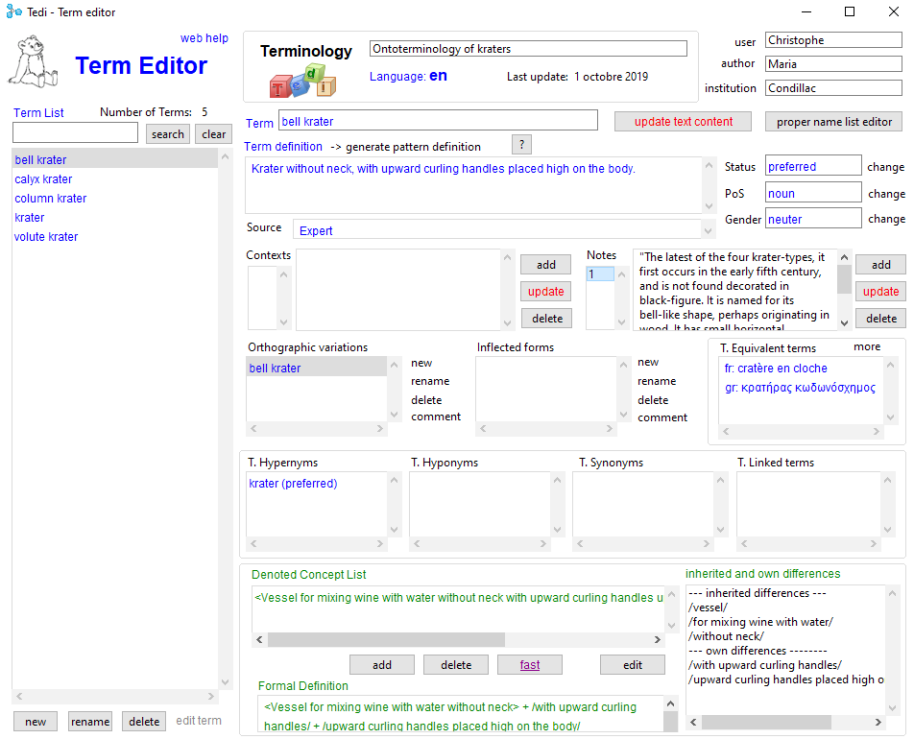


Figure 4. Éditeur de termes de Tedi

L'objectif étant de proposer des définitions de termes qui reposent sur la définition formelle des concepts dénotés par les termes, Tedi génère automatiquement des patrons de définition en langue naturelle. Ces patrons sont construits à partir du terme préférentiel dénotant le concept générique et des caractéristiques essentielles spécifiques au concept. Il reste à l'expert à éditer et/ou compléter la définition proposée par Tedi (voir figure 4).

Enfin, Tedi permet l'export des ontoterminologies aux formats RDF/OWL permettant ainsi de les importer dans Protégé et/ou de les interroger à l'aide SPARQL, CSV (pour une édition sous CmapTools par exemple), SKOS, JSON, et HTML (statique, dynamique). Un export au format HTML dynamique permet de visualiser directement l'ontoterminologie sous la forme d'un dictionnaire électronique (voir figure 5).

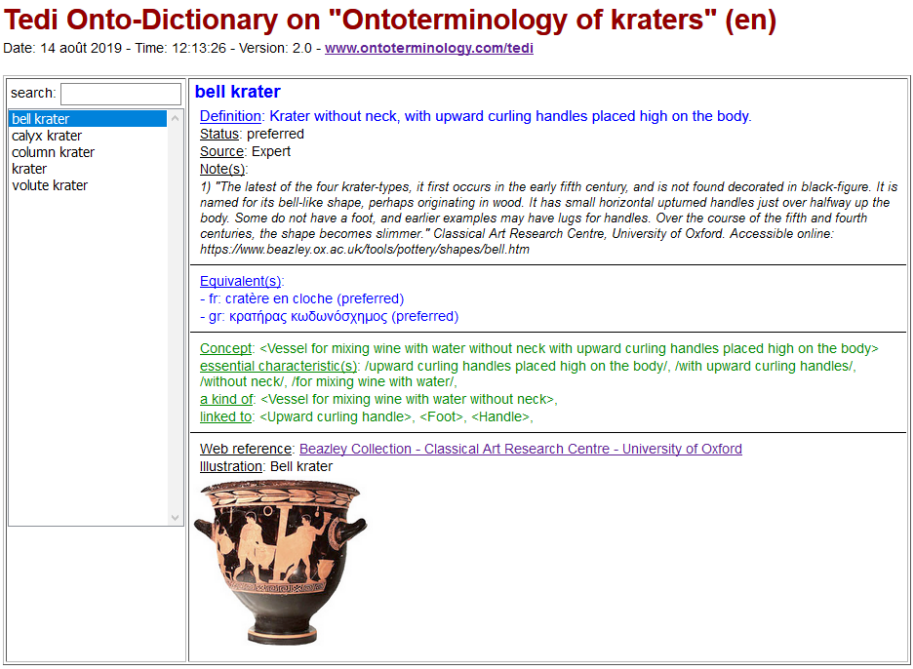


Figure 5. Export au format HTML dynamique

5. Comparaison

Les critères de comparaison pris en compte portent sur le logiciel lui-même (architecture, disponibilité, ergonomie, interopérabilité, documentation), sur la dimension conceptuelle (théorie du concept), la dimension linguistique (théorie du terme), la méthodologie (construction de la terminologie et de l'ontologie) et enfin le point de vue de l'expert (prise en main, autonomie, réponse à ses besoins).

5.1. Architecture logicielle

Protégé est implémenté en Java et fonctionne sur une large gamme de plates-formes logicielles, notamment Windows, MacOS, Linux et Unix. Il est construit en utilisant une architecture à plusieurs niveaux, fournissant une

couche de stockage ontologique, une couche de modèle de connaissances, une interface utilisateur graphique (GUI) et une interface de programmation d'application (API). En outre, Protégé possède une architecture de plug-in, permettant aux développeurs d'étendre les fonctionnalités de base de Protégé de nombreuses manières sans avoir à modifier le code source de Protégé (https://protegewiki.stanford.edu/wiki/Protege_Plugin_Library) (Noy *et al*., 2007). Protégé, permet d'importer et d'exporter des ontologies dans de nombreux formats différents, notamment RDF/XML, OWL/XML, JSON-LD, Turtle/N3, OBO, etc. Il permet l'import d'ontologies favorisant ainsi la réutilisation et l'organisation modulaire, l'accès *via* des espaces de noms à de vocabulaires disponibles sur le web.

Contrairement à Protégé, Tedi est écrit dans un langage propriétaire (Smalltalk /Visualworks) et son utilisation, gratuite à des fins de recherche et d'enseignement, est soumise à un protocole d'accord avec l'Université Savoie Mont-Blanc. Il est disponible en version « stand-alone » sur Windows et MacOs. Tedi permet d'exporter les ontoterminologies aux formats RDF/OWL, CSV, SKOS, JSON et HTML statique et dynamique. Sans aucune comparaison possible avec Protégé, Tedi offre néanmoins un certain nombre de documents : manuels, publications, vidéos. Un site web est dédié à l'ontoterminologie et à Tedi : http://ontoterminology.com/tedi

5.2. Dimension conceptuelle

Protégé permet la modélisation des entités de l'ontologie (classes et relations) dans le formalisme OWL2. Une classe est définie par des conditions nécessaires ou nécessaires et suffisantes. Dans ce second cas, il s'agit de « classe définie » permettant le raisonnement par classification. Les relations sont définies par leur signature et peuvent être caractérisées par des propriétés mathématiques dont les plus utilisées sont la symétrie et la transitivité. Les classes sont structurées selon une hiérarchie inclusive. Les intensions des propriétés sont héritées par les sous-classes.

Tedi implémente une théorie du concept dédiée à la Terminologie (Roche 2020) qui repose sur les notions de caractéristiques essentielles et de caractéristiques descriptives. Un concept est défini comme une combinaison unique de caractéristiques essentielles, les concepts se structurant en un système à travers différentes relations (générique, partitive, associative). Cette approche de la conceptualisation d'un domaine est familière aux experts tout en respectant les principes terminologiques de l'ISO telles que les prônent les normes 1087-1 et 704.

5.3. Dimension linguistique

L'interface Protégé permet si besoin de définir les termes désignant les entités de l'ontologie (concept et propriétés) sous forme de couples (label, langage). Ces métadonnées s'appuient sur des vocabulaires standards (RDF, RDFS, SKOS, etc.) ou définis par le concepteur de l'ontologie. L'utilisation de ces vocabulaires permet d'exprimer le statut du terme (préférentiel, alternatif, etc.), les définitions, les notes, etc. Il est également possible en important le module ontolex (https ://jogracia.github.io/ontolex-lexicog/) de travailler sur la dimension linguistique et faire le lien avec le concept ainsi défini.

Les termes sont explicitement représentés dans Tedi, ils ne sont pas réduits à de simples étiquettes sur des concepts. La dimension linguistique a autant d'importance que la dimension conceptuelle. Un éditeur spécifique lui est dédié (figure 4). Il permet de préciser le statut du terme (préférentiel, alternatif, etc.), les contextes, les notes, les formes fléchies mais aussi de calculer les synonymes, hyperonymes et hyponymes terminologiques, ainsi que de générer des patrons de définition du terme sur la base de la définition formelle du concept dénoté par le terme.

5.4. Méthodologie de construction

Protégé est un outil très complet d'édition d'ontologies et est indépendant de toutes méthodologies de construction.

Tedi implémente une méthode de construction d'ontologies guidée par le terme, un concept étant un ensemble de caractéristiques suffisamment stable pour être nommé en langue. Tedi guide également l'expert tout au long de sa construction de l'ontologie et de la terminologie en ne proposant que ce qui est valide à un moment donné, par exemple, les caractéristiques essentielles encore disponibles pour la définition d'un nouveau concept ou les concepts génériques possibles. La vérification des propriétés logiques est faite « à la volée » et non *a posteriori*.

5.5. Le point de vue de l'expert

Protégé et Tedi sont deux environnements relativement difficiles à appréhender. Les interfaces sont riches, voire complexes. L'utilisation de Protégé peut difficilement se faire sans l'aide d'un ingénieur de la connaissance et ce malgré une documentation très riche. On regrette également que la dimension linguistique soit réduite à des annotations et qu'il n'y ait pas d'interfaces qui

lui soient dédiées. Enfin, même si on se doute que Protégé possède un champ d'applications beaucoup plus vaste que Tedi, celui-ci, du fait même qu'il soit dédié à une seule tâche, la construction d'ontoterminologies multilingues, et qu'il intègre une méthode de construction, répond davantage aux besoins de l'expert. Après une phase d'adaptation, on apprécie les interfaces dédiées, les nombreuses aides à la construction de l'ontologie et de la terminologie en ne proposant que ce qui est valide. On apprécie également la génération automatique de patrons de définition en langue naturelle sur la base de la définition formelle du concept et la possibilité de générer un HTML dynamique qui permet de visualiser directement l'ontoterminologie qu'on vient de créer sous la forme d'un dictionnaire électronique.

6. Conclusion

Protégé accessible sous une licence open source est le logiciel le plus utilisé pour la construction et la maintenance d'ontologies [Musen, 2015]. Il applique les recommandations du W3C et il est supporté par une large communauté d'utilisateurs contribuant avec l'équipe Protégé à l'améliorer. Il permet d'inspecter, de parcourir, de codifier, de modifier les ontologies et de permettre de cette manière leur développement et leur maintien. Il existe actuellement une version complète standalone (Protégé 5) dotée de nombreux plugins et une version allégée collaborative (Web protégé) offrant un accès distribué à une ressource ontologique à l'aide de n'importe quel navigateur Web. Enfin, *Protégé a une visée universelle au sens où il ne se limite pas à la construction d'ontologies à des fins terminologiques.*

Contrairement à Protégé qui a débuté dans les années 80, Tedi est un logiciel récent (la version 1 date de 2018) et n'a pas une visée universelle. *Il est destiné aux experts pour la construction de terminologies en respectant les principes terminologiques des normes ISO.* Tedi est écrit dans un langage qui n'est pas libre (Smalltalk) et son utilisation est soumise à accord préalable. Les avantages de Tedi reposent, d'une part, sur une théorie du concept et du terme proche de la compréhension qu'en ont les experts (définition aristotélicienne en genre prochain et différence spécifique), et, d'autre part, sur l'implémentation d'une méthodologie de construction qui guide l'expert du domaine dans la définition du concept et dans sa dénomination tout comme dans l'écriture de définition en langue naturelle ; avantages qui permettent à l'expert d'être autonome.

Protégé et Tedi diffèrent sur de nombreux points qu'on pourrait résumer en disant que le premier est aussi universel que le second est spécifique. Le choix entre l'un ou l'autre se fera principalement au regard de la finalité du projet pour laquelle on est amené à construire une terminologie et une ontologie.

Bibliographie

AAT Getty (2019), http://www.getty.edu/research/tools/vocabularies/aat/ « amphora » http://www.getty.edu/vow/AATServlet?english=N&find=amphora&logic=AND&page=1&note, consultation le 30 octobre 2019.

Azzi, R., Despres, S. and Nobecourt, J. (2019) « MCVGraphViz : a Web Tool for Knowledge Dynamic Visualization ». RIVF 2019 - The 2019 IEEE-RIVF International Conference on Computing and Communication Technologies - Danang, Vietnam, March 20-22, 2019 (track computer-science)

Bachimont B. (2000). Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingénierie des connaissances. In Z. M. Charlet J., Kassel G., Bourigault D., (Ed.), *Ingénierie des connaissances. Évolution Récentes et nouveaux défis* (p. 305-323). Paris : Eyrolles.

Beazley (2019) Classical Art Research Centre, Beazley Archive, University of Oxford. https ://www.beazley.ox.ac.uk/tools/pottery/shapes/default.htm. Consulté le 30 octobre 2019.

Boardman, J. (1974) *Athenian Black Figure Vases*. Thames and Hudson. ISBN 0500201382 (chapitre 9 Shapes, names, and functions pp. 184-192)

Boardman, J. (1975) *Athenian Red Figure Vases : The Archaic Period*. Penguin. (chapitre 5 Shapes and dates pp. 208-211)

Cartwright, M. (2013). *A Visual Glossary of Greek Pottery*. Ancient History Encyclopedia. Retrieved from https ://www.ancient.eu/article/489/ on 30th October 2019

Cook R. M. (1997) *Greek Painted Pottery*. Routledge, ISBN 0415138590, 9780415138598 (chapitre 8 Shapes pp. 207-230)

Dandan, R., Desprès, S. and Nobécourt, J. (2018) OAFE : An Ontology for the description of elderly activities. SITIS 2018 - The 14th International Conference on SIGNAL IMAGE TECHNOLOGY & INTERNET BASED SYSTEMS - 26-29 November 2018 - Las Palmas de Gran Canaria, Spain (track Whorkshop KARE)

Despres S. (2016) « Construction d'une ontologie modulaire. Application au domaine de la cuisine numérique. Revue d'Intelligence Artificielle », 30(5): 509-532 (2016).

Despres S. (2018) sensoMIAM Le module sensoriel de l'ontologie MIAM. In JFO 2018 - 7e Journées Francophones sur les Ontologies - 1,2 novembre 2018, Hammamet, Tunisie (papier long ; http://www.jf-ontologies.net/index.php/papiers-acceptees)

Gruber, Thomas (1992) « A Translation Approach to Portable Ontology Specifications » Knowledge Acquisition 5 (2), p. 199-220

Guarino, Nicolas ; Oberle, Daniel ; Staab, Steffen (2009), « What Is an Ontology » in Staab, Steffen ; Studer, Rudi (dir.) Handbook on Ontologies, Springer-Verlag, Berlin, ISBN 978-3-540-92673-3, p. 1-16.

Horridge M., Tudorache T., Vendetti J., Nyulas C., Musen M., Noy N. (2013). Simplified OWL ontology editing for the web : is WebProtégé enough ? *The Semantic Web - ISWC 2013, Proceedings part I - 12th International Semantic Web Conference*, Sydney New South Wales, Australia, p. 200-215.

ISO 1087-1. (2000). Terminology work - Vocabulary - Part 1 : Theory and application. Geneva : International Standards Organisation

Musen, M.A. and Team Protégé (2015) *The Protégé project : A look back and a look forward*. AI Matters. Association of Computing Machinery Specific Interest Group in Artificial Intelligence, 1(4), June 2015. DOI :10.1145/2557001.25757003.

Richter, Gisela and Marjorie Milne (1935) Shapes and Names of Athenian Vases https ://libmma.contentdm.oclc.org/digital/collection/p15324coll10/id/146188

Roche, C. and Papadopoulou, M. (2019), « Mind the Gap : Ontology Authoring for Humanists », 1st International Workshop for Digital Humanities and their Social Analysis (WODHSA)- Episode V : The Styrian Autumn of Ontology, September 23-25, a Workshop hosted by Joint Ontology Workshops, Medical University of Graz (Austria), September 23-25, 2019

Papadopoulou, M. and Roche, C. (2018), « Ontologization of Terminology. A worked example from the domain of ancient Greek dress », AIDAinformazioni Journal, v. XXXVI, n° 1-2, p. 89-107

Roche, C. Papadopoulou, M. (2019), « Terminologie et Ontologie pour les Humanités Numériques : Le cas des vêtements de la Grèce antique », Revue Humanités Numériques, n° 2, 2019

Roche, C. Papadopoulou, M. (2018), « Définition ontologique du terme. Le cas des vêtements de la Grèce antique », LTT 2018, Lexicologie Terminologie Traduction, Grenoble (France), 27-28 septembre 2018

Roche, C. (2005), « Terminologie et ontologie », Revue Langages, Larousse, 39e année, n° 157, p. 48-62

Roche, C. (2007), « Le terme et le concept : fondements d'une ontoterminologie », in Roche, Christophe (dir.) *Terminology & Ontology : Theories and application, TOTh 2007*, France, p. 1-22

Roche, C. (2020, à paraître) *Une Théorie du Concept pour la Terminologie*, Presses Universitaires Savoie Mont Blanc, coll. Terminologica, Chambéry.

Rubin L.R., Noy N.F., Musen M. (2007), Protégé : A Tool for Managing and Using Terminology in Radiology Applications, Journal of Digital Imaging, 20(Suppl 1), p. 34-46

Schreiber, T. (1999). Athenian Vase Construction. A Potter's Analysis, J. Paul Getty Museum (Part II et Appendixes)

Studer R., Benjamins V. R, Fensel D. (1998). D. Knowledge Engineering : Principles and methods. Data & Knowledge Engineering 25, p. 161-197

Suárez-Figueroa, M.C., Gomez-Perez, A., (2008). Towards a Glossary of Activities in the Ontology Engineering Field, LREC

Suarez-Figueroa M.C., Gomez-Perez A., Fernandez-Lopez M. (2015). The NeOn Methodology framework : A scenario-based methodology for ontology development. Applied Ontology 10(2), p. 107-145

Diaterm : un modèle pour représenter l'évolution diachronique des terminologies dans le web sémantique

Silvia Piccini*, Andrea Bellandi*, Matteo Abrate**

* Istituto di Linguistica Computazionale Antonio Zampolli - CNR
Via Moruzzi 1, 56124 Pisa
silvia.piccini@ilc.cnr.it
andrea.bellandi@ilc.cnr.it
**Istituto di Informatica e Telematica - CNR
Via Moruzzi 1, 56124 Pisa
matteo.abrate@iit.cnr.it

Abstract. La présente contribution vise à illustrer *Diaterm*, un modèle à trois niveaux consacré à la représentation formelle de l'évolution diachronique des concepts et des termes dans un domaine donné. L'approche adoptée est basée sur la réification des relations N-aires. De surcroît, un ensemble de règles SWRL a été conçu permettant d'effectuer des tâches de raisonnement et d'automatiser l'attribution des informations temporelles. *Diaterm* a été adopté pour représenter formellement l'évolution diachronique de la terminologie de Ferdinand de Saussure telle qu'elle émerge dans les œuvres du Maître genevois.

La science cherche le mouvement perpétuel. Elle l'a trouvé ; c'est elle-même. La science est continuellement mouvante dans son bienfait. Tout remue en elle, tout change, tout fait peau neuve. Tout nie tout, tout détruit tout, tout crée tout, tout remplace tout. Ce qu'on acceptait hier est remis à la meule aujourd'hui. La colossale machine Science ne se repose jamais ; elle n'est jamais satisfaite ; elle est insatiable du mieux, que l'absolu ignore.

Victor Hugo

1. Introduction

La dimension diachronique a longtemps été le parent pauvre de la terminologie, les études consacrées à l'histoire des termes et des concepts étant sporadiques et se concentrant presque toutes dans les vingt dernières années (par exemple van Campenhoudt 1997, Humbley 2011[1], Zanola 2014, Piccini 2016).

Cette résistance aux aspects historico-diachroniques est essentiellement liée à une vision conceptuelle et universaliste à la base de la théorie « classique » de la terminologie, inaugurée par Eugen Wüster (1968). Le terme – fondamentalement monosémique – est conçu comme une étiquette associée à un concept – essentiellement unidimensionnel – d'une manière immuable et indifférente à toute variation culturelle et temporelle. En d'autres termes, les structures conceptuelles sont universelles et statiques ainsi que leurs termes et leurs référents.

Dans un état de science normale, en fait, lorsque le développement des connaissances adhère à un paradigme (Kuhn 1970), le système terminologique est standard, stable, défini explicitement et utilisé de manière consensuelle par la communauté scientifique de référence.

Le terme s'éloigne de sa nature de signe linguistique pour devenir de plus en plus équivalent à un symbole, à un « désignateur rigide » (Kripke 1995), évoquant à certains égards la conception néo-positiviste référentielle du signe à la base de l'onomantique de Riggs. Un certain contrôle social est exercé sur le sens des termes pour garantir l'absence d'une « ambiguïté ponctuelle » (Thoiron et Béjoint 2010, 108) et une communication efficace entre les spécialistes d'un domaine.

On parle dans ces contextes de terminologies paradigmatiques (Cosenza 2016), qui s'opposent aux terminologies théoriques caractérisées par une grande instabilité typique des phases de science révolutionnaires. Dans les périodes précédant ou suivant un état de « science normale », lorsque les idées qui constituent le paradigme sont remises en question, on assiste souvent à des fluctuations terminologiques, la frontière entre mot et terme devenant plus fluide, plus poreuse, plus perméable.

1 L'auteur, dans son article dédié à la néologie rétrospective, donne un aperçu exhaustif des études terminologiques – conduites dans le contexte français – qui prennent en compte les aspects diachroniques.

Pour représenter formellement l'évolution diachronique des concepts et des termes dans ces contextes, afin que cette formalisation soit exploitable par l'ordinateur, un modèle a été conçu, nommé *Diaterm*, faisant l'objet du présent article.

2. La terminologie historico-diachronique

Les systèmes notionnels et les terminologies sont en constante évolution comme le démontre l'histoire de la science.

Au fil des siècles, les savants ont élaboré différents modèles pour répondre aux nouvelles exigences théoriques dérivant de l'observation, donnant lieu parfois à une véritable révolution scientifique. Révolutionner signifie abandonner l'ancien paradigme et par conséquent changer l'ancienne terminologie : de nouveaux termes sont introduits pour exprimer le nouveau système de concepts et des termes anciens sont soumis à un changement sémantique ou rejetés lorsque le concept devient obsolète.

On comprend bien que dans ces contextes le texte – sous forme de corpus – acquiert une valeur fondamentale, car il permet de tracer et reconstruire l'évolution du système terminologique d'un domaine donné. Les phénomènes de néosémie et néonymie sont documentés dans la littérature ou – de manière encore plus emblématique – dans les pages intimes des savants, dans leurs notes, dans leurs brouillons ou dans la correspondance échangée avec des collègues sur des sujets scientifiques.

Le texte constitue pour le savant le lieu où il pense, réfléchit et organise ses idées. Il s'agit d'un espace stratégique où les idées prennent forme, pour trouver finalement leur formulation linguistique, leur terme.

La terminologie historico-diachronique peut être donc définie – à juste titre – une terminologie textuelle, une terminologie du discours (Bourigault et Slodzian 1999).

La démarche du terminologue devient sémasiologique, le point de départ étant constitué par le sens d'un terme dans un contexte linguistique réel. L'accent est mis sur la langue telle qu'elle est réellement utilisée (*Ist-Norm*), dans de véritables textes spécialisés et non sur la langue telle qu'elle devrait être utilisée (*Soll-Norm*). Le terme révèle ainsi sa nature de signe fonctionnant dans un système linguistique spécifique à une société, une culture, une époque donnée.

On le verra dans le cas emblématique consacré à la figure de Ferdinand de Saussure et discuté dans la Section 5.

Bien qu'il s'agisse d'une diachronie courte, le corpus saussurien permet d'observer *in vivo* et retracer le « processus de naissance conceptuelle et d'invention théorique » (Fenoglio 2012, 13) qui passe inéluctablement à travers des réécritures continues du texte manuscrit et des changements linguistiques perpétuels à la recherche du terme le plus adéquat.

3. Modéliser la diachronie dans le web sémantique

Le but du présent travail était de construire un modèle pour représenter formellement l'évolution diachronique de concepts et de termes dans un domaine donné, le seul impératif méthodologique étant l'adoption des technologies du web sémantique, telles que RDF, OWL, SPARQL et SWRL, afin de produire – en conformité avec la philosophie de la science ouverte – des données FAIR (Wilkinson 2016), à savoir faciles à trouver, accessibles, interopérables et réutilisables par l'homme et la machine.

Représenter l'évolution diachronique de l'information en OWL constitue, toutefois, un véritable défi.

Le principal problème est représenté par le fait qu'OWL, le langage standard pour la représentation et le partage des ontologies dans le web sémantique, constitue un langage monotonique. Cela équivaut à dire que les nouvelles informations introduites ne peuvent pas désavouer la valeur de vérité des informations précédemment formalisées. Par conséquent, des scénarios statiques seuls peuvent être représentés, l'ajout d'une nouvelle dimension, celle temporelle dans ce contexte, posant d'énormes problèmes de calculabilité et de décidabilité.

Pour surmonter ces limites, différentes approches ont été proposées en littérature telles que le *versioning* (Klein and Fensel 2001), le modèle *perdurantiste* ou *4-D fluents* (Welty *et al*. 2006) ou encore le modèle N-aire.

Notre choix s'est porté sur le modèle N-aire pour deux raisons principales.

Tout d'abord, l'approche des relations N-aires est fortement recommandée par le groupe de travail *Ontology Engineering and Patterns Task Force of the Semantic Web Best Practices and Deployment Working Group*[2].

2 https://www.w3.org/TR/swbp-n-aryRelations/

Ensuite, le modèle N-aire nécessite l'introduction d'un nombre limité d'objets par rapport à l'approche perdurantiste et donc « outperforms the 4D-fluents representation in terms of required assertions (axioms) and consequently in reasoning time » (Batsakis *et al*. 2017, 14).

Plus spécifiquement, les relations d'arité supérieure à 2 sont exprimées à travers le premier patron de conception d'ontologies *(ODP1 Ontology Design Pattern*[3]) proposé par le W3C (Noy *et al*. 2006), qui se base sur le processus appelé « réification ». Réifier une relation d'arité supérieure à 2 signifie la représenter en tant que classe.

Il faut toutefois remarquer que l'approche N-aire – comme l'approche perdurantiste – tout en assurant la décidabilité, est définie « invasive » car elle modifie les définitions des classes ou des relations qui doivent être « temporalisées », entraînant ainsi l'invalidation des mécanismes de raisonnement.

Considérons par exemple le triplet *hypernym(a, b)* valide dans t_1 et *hypernym(b, c)* valide dans t_2 – *a*,*b*,*c* étant des termes caractérisés par une acception spécifique et t_1, t_2 deux différents intervalles temporaux. La synonymie étant définie comme une propriété transitive, le moteur d'inférence devrait inférer le triplet suivant : *hypernym(a, c,* $t_1 \cap t_2$*)*, à savoir que les termes *a* et *c* sont synonymes dans un intervalle temporel correspondant à l'intersection de t_1 et t_2. Le moteur d'inférence par contre infère le triplet *hypernym(a, c)*.

Comme on le verra dans la Section suivante, pour surmonter cette limite inévitable avec l'adoption du modèle N-aire, un ensemble de règles a été proposé reproduisant les mécanismes de raisonnement et prenant en considération la validité temporelle des relations.

4. Le modèle *Diaterm*

Le modèle *Diaterm* (Figure 1) est caractérisé par une architecture à trois niveaux : un niveau ontologique/conceptuel, un niveau terminologique/linguistique et un niveau textuel.

L'axe diachronique est transversal à tous ces plans, c'est-à-dire que le changement peut se vérifier dans les trois ordres. Des modifications, en effet, peuvent affecter le texte, qui souvent – dans sa phase de gestation avant d'arriver à la presse – est soumis à une réécriture continue ; ou encore ce sont le plan terminologique et/ou le plan conceptuel qui peuvent subir des variations.

3 http://www.w3.org/2001/sw/BestPractices

La séparation du plan linguistique et du plan conceptuel – théoriquement assez débattue – se révèle très utile d'un point de vue méthodologique, permettant ainsi de distinguer clairement quand une notion technique demeure identique, tout en changeant sa dénomination, ou quand, au contraire, un changement terminologique correspond à un changement significatif dans la théorie.

4.1. Le niveau conceptuel

Dans le plan conceptuel, les concepts désignés par les termes ainsi que leur évolution reçoivent une description formelle sous forme d'une ontologie.

La réification étant le patron de conception d'ontologies adopté dans *Diaterm*, la classe dcmitype :Event joue un rôle central. Elle représente la réification d'une propriété qui subsiste dans un intervalle de temps donné *t*[4].

De surcroît, on a introduit la propriété d'objet during[5], ayant comme domaine et comme co-domaine respectivement la classe dcmitype :Event et la classe time :ProperInterval[6] tirée de l'ontologie OWL-Time et la propriété d'objet time :temporalProperty, qui relie les instances de la classe dcmitype :Event respectivement aux individus source et cible de la propriété d'origine réifiée.

Comme on le verra successivement, les propriétés d'objet converties en propriétés temporelles ont été formalisées en tant que sous-propriétés de time :temporalProperty.

4.2. Le niveau terminologique

Afin de représenter les informations linguistiques concernant les termes et leur évolution diachronique, le model *OntoLex-Lemon* (McCrae *et al*. 2017) a été adopté et modifié selon les principes de la réification. Par conséquent,

4 En accord avec la philosophie des données ouvertes et liées, la classe Event a été empruntée au type dcmitype :Event tiré de la Dublin Core Metadata Initiative (DCMI) et défini comme une « occurrence non persistante, basée sur le temps » (cf. http://purl.org/dc/dcmitype/Event).

5 La propriété during a été modélisée en tant que sous-propriété du type dcterms :date, à savoir « heure ou moment de l'heure associée à l'événement dans le cycle de vie de la ressource ».

6 Cette classe est définie dans la DCMI comme « une entité temporelle avec une étendue ou une durée non nulle, dont les valeurs de début et de fin sont différentes » (cf. https ://www.w3.org/TR/owl-time/#time :ProperInterval).

toutes les relations prévues dans le modèle ont été converties en sous-classes de la classe Event.

On a donc introduit : la classe senseEvent, qui constitue la réification de la relation ontolex :sense reliant un individu de la classe ontolex :LexicalEntry à un sens spécifique ; la classe ReferenceEvent, représentant la réification de la relation ontolex :reference entre le sens et le concept ontologique ; les classes EquivalentEvent, IncompatibleEvent, BroaderEvent et NarrowerEvent, qui représentent dans l'ordre la réification des relations *OntoLex-Lemon* lexinfo :synonym (synonymie), lexinfo :antonym (antonymie), lexinfo :hypernym (hyperonymie) et lexinfo :hyponym (hyponymie)[7].

La classe Event – ainsi que ses sous-classes – est reliée à la classe time :ProperInterval à travers la propriété d'objet during, qui indique l'intervalle de temps dans lequel se produit l'événement en question.

En particulier, les relations de *OntoLex-Lemon* étant réifiées et transformées en classes, six nouvelles relations lexico-sémantiques ont été introduites comme sous-propriétés de time :temporalProperty, à savoir temporalSense, temporalReference, temporalEquivalent, temporalIncompatible, temporalNarrow, et temporalBroader. Elles font référence à la validité temporelle de leurs événements respectifs et relient les instances de la partie temporelle de l'ontologie.

Ainsi, par exemple, la classe EquivalentEvent est liée à la classe time :ProperInterval à travers la relation temporalEquivalent, en spécifiant la période dans laquelle deux sens sont utilisés comme synonymes dans les différents contextes.

4.3. Le niveau textuel

Étant donné que la terminologie historico-diachronique est par définition une terminologie textuelle, dans *Diaterm* le concept d'attestation joue un rôle central. Plus précisément, le sens d'un terme est lié au texte où il est attesté, et pour chaque texte la période d'écriture ou de publication est définie au moyen de relations spécifiques.

Dans le niveau textuel de *Diaterm* les entités suivantes ont été introduites :

- la classe dcmitype :Text, qui représente tout produit de l'activité intellectuelle consistant en une « réalisation discursive d'un système de

7 LexInfo est une ontologie qui a été définie pour fournir des catégories linguistiques pour le modèle OntoLex-Lemon (https://lexinfo.net/).

signes ou d'un système de significations » (TLFI). Le terme « texte » doit être doc interprété ici au sens large, incluant des entités très différentes les unes des autres, telles que des livres, des articles publiés, des notes manuscrites, des brouillons d'œuvres inachevées, etc[8]. ;

- la propriété d'objet lawd :hasAttestation, qui relie un sens, instance de la classe ontolex :LexicalSense, à une instance de la classe dcmitype :Text[9] ;
- la propriété d'objet hasWritingTime, définie comme une sous-propriété de dcterms :date, qui relie un individu de la classe dcmitype :Text à un individu appartenant à la classe time :ProperInterval de l'ontologie OWL-Time[10]. En ce qui concerne les textes publiés, au lieu de la propriété hasWritingTime, la relation de la DCMI dcterms :issued peut être utilisée, l'année de publication constituant la fin de la période de rédaction.

8 Il peut être considéré comme équivalent du type dcmitype :Text tel qu'il est défini par la DCMI, à savoir une « ressource dont le contenu est principalement constitué de mots à lire » (cf. http://purl.org/dc/dcmitype/Text). Les spécialistes intéressés à des distinctions plus subtiles peuvent recourir à d'autres modèles tels que FRBR ou Bibframe. Il est important de souligner que si un vocabulaire différent est adopté, la validité du modèle ici proposé n'est pas compromise.

9 Dans le but de réutiliser les vocabulaires existants, la propriété hasAttestation a été adoptée, telle qu'elle est définie dans l'ontologie LAWD (*Linked Ancient World Data*). Cf. http://lawd.info/ontology/hasAttestation.

10 Dans le cas où l'expert n'est pas en mesure de définir avec précision la période pendant laquelle un auteur a travaillé à la rédaction d'un texte manuscrit, il peut identifier – à l'aide d'éléments internes au texte ou de sources externes – des intervalles qualitatifs (« avant X », « après Y », etc.), à savoir les limites d'un intervalle temporel dans lequel le manuscrit a certainement été écrit.

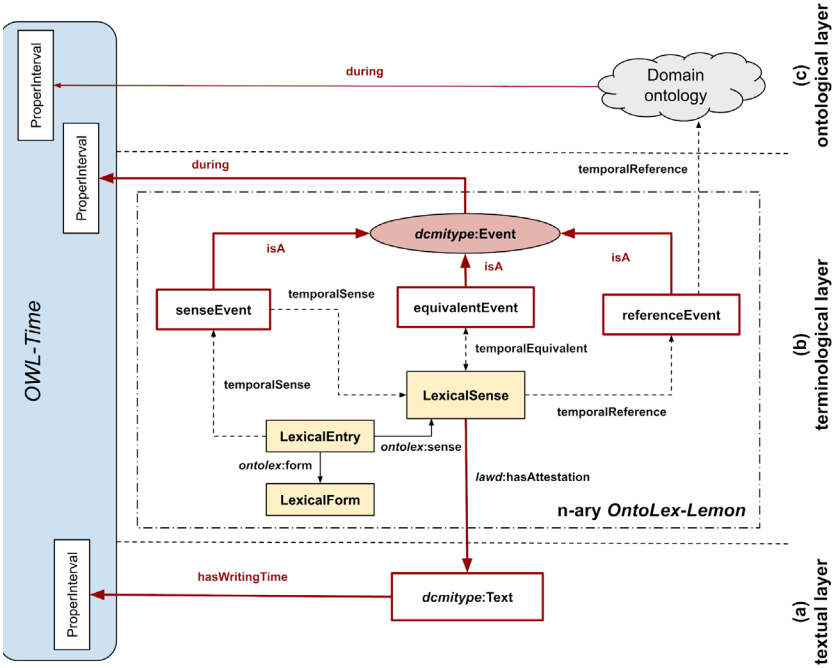


FIG. 1 – *Le modèle Diaterm : (a) le niveau textuel ; (b) le niveau terminologique reposant sur une version N-aire du modèle OntoLex-Lemon ; (c) le niveau ontologique constitué par les définitions formelles en OWL des concepts désignés par les termes.*

4.4. Le mécanisme de raisonnement dans *Diaterm*

L'adoption d'OWL en tant que langage de représentation permet d'utiliser un ensemble des règles SWRL pour automatiser l'attribution de certaines informations temporelles, comme la période d'attestation d'un terme ou d'un concept spécifique.

La tâche de raisonnement est réalisée à travers ces règles SWRL qui ont été créées *ad hoc* et qui utilisent les règles d'Allen[11].

Voici un exemple d'une règle qui permet d'attribuer automatiquement la validité temporelle d'un sense à partir de son attestation dans un texte qui a été composé dans un intervalle de temps donné :

> dcmitype :Text(?t) ^ time :ProperInterval(?i) ^ hasWritingTime(?t, ?i) ^ ontolex :LexicalEntry(?l) ^ ontolex :LexicalSense(?s) ^ isAttestedIn(?s, ?t) ^ ontolex :sense(?l, ?s) ^ swrlx :createOWLThing(?se, ?s) → SenseEvent(?se) ^ temporalSense(?l, ?se) ^ temporalSense(?se, ?s) ^ during(?se, ?i)

Si le sens d'une entrée lexicale est attesté dans plusieurs textes – chaque texte étant écrit dans un intervalle temporel spécifique – , la règle sera activée pour chaque texte, reliant ainsi l'entrée lexicale à ce sens dans une période plus complexe constituée d'intervalles de temps multiples.

D'autres règles ont été élaborées permettant de définir par exemple la validité temporelle d'une certaine notion, ou de créer les équivalents diachroniques des relations de synonymie, d'antonymie, d'hyperonymie et d'hyponymie.

5. Interrogation et visualisation : le cas de Saussure

Le modèle *Diaterm* a été adopté pour représenter formellement l'évolution diachronique de la terminologie du linguiste genevois Ferdinand de Saussure, telle qu'elle émerge du corpus textuel (1874-1911).

L'un des principaux avantages de cette complexe formalisation est la possibilité de reconstruire, à travers de nombreuses requêtes sophistiquées et complexes, les dynamiques évolutives qui ont caractérisé le système terminologique et conceptuel tel qu'il est attesté dans un corpus.

Afin d'aider l'utilisateur à interpréter les données contenues dans la ressource termino-ontologique diachronique, une visualisation a été réalisée permettant de comparer les différents états synchroniques d'un système conceptuel et terminologique, ces états se succédant dans le temps.

Comme on le voit dans la Figure 2, la ressource reçoit une représentation arborescente. En appuyant sur une icône en haut à gauche, il est possible de

11 Dans la théorie d'Allen, 13 relations de base (ou relations atomiques) décrivent toutes les manières possibles d'ordonner les extrémités de deux intervalles.

sélectionner des intervalles différents. Afin d'aider l'utilisateur dans la comparaison, la représentation de chaque intervalle conserve une trace semi-transparente des nœuds et des relations qui caractérisent les autres intervalles.

La technique de visualisation adoptée consiste en une réadaptation du layout proposé par Bezerianos *et al.* (2010) pour la représentation des arbres généalogiques, la structure des données de ces derniers présentant des caractéristiques similaires aux relations taxonomiques d'une ressource terminologique. Toutes les deux sont, en effet, des graphes acycliques (*Directed Acyclic Graphs* - DAG), qui présentent de fortes similitudes avec une structure arborescente. La présence de plusieurs parents pour certains nœuds empêche l'utilisation de techniques adoptées normalement pour représenter des structures hiérarchiques. Cependant, l'utilisation d'un layout approprié pour les graphes ou pour les graphes acycliques rend le diagramme moins lisible ou ne code pas correctement l'aspect hiérarchique.

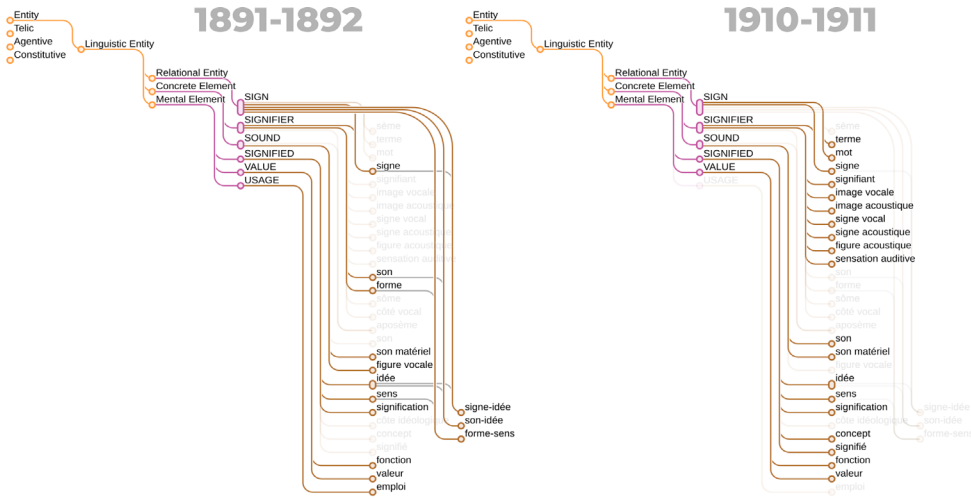


FIG. 2 – *Visualisation de la terminologie consacrée à la sémiologie chez Saussure en deux intervalles de temps (1891-1892 et 1910-1911) : en jaune les classes et les relations de subsomption, en violet les instances et les relations d'instanciation, en marron les termes et la relation de référence. La relation de synonymie n'est pas explicitement tracée mais déductible du fait que les termes dénotant le même concept se trouvent adjacents les uns aux autres dans les derniers niveaux de la hiérarchie.*

Dans la figure 2 on présente une photographie de la terminologie consacrée à la réflexion sémiologique du maître genevois telle qu'elle est attestée dans *De la double essence du langage* (datant plausiblement de 1891-1892) et dans le *Troisième cours de linguistique générale* (1910-1911).

La comparaison entre les deux états synchroniques met en relief comme les trois néologismes introduits par Saussure en 1891-1892 pour dénoter le signe en tant qu'unité dyadique (*signe-idée*, *son-idée*, *forme-sens*) ont été tous abandonnés successivement. L'ambiguïté engendrée par la polysémie du terme *signe* est éliminée seulement en 1911, lorsque Saussure propose de conserver le terme pour désigner l'entité linguistique dans son ensemble, et remplacer ainsi les termes *signe* en tant que partie matérielle du signe linguistique *(forme)* et *concept* respectivement par *signifiant* et *signifié*.

6. Conclusions et nouvelles perspectives de recherche

Les potentialités offertes par l'adoption de *Diaterm* sont énormes, comme le cas de Ferdinand de Saussure le démontre. Ce modèle présente, toutefois, des points faibles : premièrement un nombre considérable de règles doivent être introduites, la plupart desquelles atteint un haut degré de complexité ; deuxièmement le fait que dans *Diaterm* les mécanismes de raisonnement sont basés exclusivement sur les règles entraîne une performance réduite en termes de temps nécessaire pour le calcul des inférences. De surcroît, cette performance diminue lorsque la taille des ressources terminologiques augmente.

L'objectif futur est donc de rendre le modèle proposé plus efficace du point de vue du calcul d'inférences, en évaluant l'adoption d'approches d'apprentissage en profondeur (*deep learning*) ou d'une représentation des ontologies à travers des bases de données relationnelles caractérisées par des index appropriés.

Références

Batsakis, Sotiris, Petrakis, Euripides G.M.., Tachmazidis, Ilias, Grigoris, Antoniou. 2017. "Temporal representation and reasoning in OWL 2." *Semantic Web* 8 (6) : 981-1000.

Bezerianos, Anastasia, Dragicevic, Pierre, Fekete, Jean-Daniel, Bae, Juhee, Watson, Ben. 2010. "GeneaQuilts : A System for Exploring Large Genealogies." In *IEEE Transactions on Visualization and Computer*

Graphics, Institute of Electrical and Electronics Engineers 16 (6), 1073-1081.

Bourigault, Didier, Slodzian, Monique. 1999. “Pour une terminologie textuelle.” *Terminologies nouvelles* 19 : 29-32.

Cosenza, Giuseppe. 2016. D*alle parole ai termini. I percorsi di pensiero di F. de Saussure*. Alessandria : Edizioni dell’Orso. Collezione “Studi e Ricerche”.

Fenoglio, Irène. 2012. “Genèse du geste linguistique : une complexité heuristique.” *Genesis* 35 : 13-40.

Humbley John. 2011. “Vers une méthode de terminologie rétrospective.” *Langages* 2011/3 (n° 183) : 51-62.

Kuhn, Thomas S. 1970. *The structure of scientific revolutions*. 2nd ed. Chicago : The University of Chicago Press.

Klein, Michel C. A., Fensel, Dieter. 2001. “Ontology versioning on the Semantic Web.” In *Proceedings of the First International Conference on Semantic Web Working*, 75-91. California : CEUR-WS.

Kripke, Saul A. 1995. *La logique des noms propres*. Paris : Les Éditions de Minuit.

Noy, Natasha, Rector, Alan, Hayes, Pat, Welty, Chris. 2006. “Defining N-ary Relations on the Semantic Web.” W3C Working Group Note, April 2006.

McCrae, John P., Bosque-Gil, Julia, Gracia, Jorge, Buitelaar, Paul, Cimiano, Philipp. 2017. “The Ontolex-Lemon model : development and applications.” In *Proceedings of eLex 2017 conference*, Leiden, The Netherlands, 19-21.

Piccini, Silvia, Bellandi, Andrea, Benotto, Giulia. 2016. “Formalizing and Querying a Diachronic Termino-Ontological Resource : the CLAVIUS Case Study.” In *Proceedings of From Digitization to Knowledge 2016 workshop* (D2K), Krakow (Poland), 38-41.

Thoiron Philippe, Béjoint, Henri. 2010. “La terminologie, une question de termes ?” *Meta* 55(1): 105-118.

Van Campenhoudt, Marc 1997. “Maille ou maillon : quand des terminographes négligent l’évolution de l’usage.” In *Proceedings of the 5th scientific days Réseau : Lexicologie, Terminologie, Traduction* (Agence Universitaire de la Francophonie), *La mémoire des mots*, edited by André Clas, Salah Mejri, Taïeb Baccouche, 251-272.

Welty, Christopher A., Fikes, Richard, Makarios, Selene. 2006. “A reusable ontology for fluents in OWL.” In *Formal Ontology in Information Systems, Proceedings of the Fourth International Conference*, FOIS 2006, Baltimore, Maryland, USA, vol. 150, 226-236.

Wilkinson, Mark D. *et al*. 2016. "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific data* 3, Vol. 3, 160018.
Wüster, Eugen. 1968. *The Machine Tool. An Interlingual dictionary of basic concepts*. 1 ed. London : Technical Press.
Zanola, Maria Teresa. 2014. *Arts et métiers au XVIII^e siècle. Essais de terminologie diachronique*. Paris : L'Harmattan.

Abstract

This paper illustrates *Diaterm*, a three-level model devoted to formally representing the diachronic evolution of concepts and terms in a given domain. The approach adopted is based on the reification of N-ary relations. A set of SWRL rules was introduced to perform reasoning tasks and automatically assign temporal information. *Diaterm* was adopted to modelling the diachronic evolution of of Ferdinand de Saussure's terminology as it emerges in the Genevan linguist's corpus.

Application of topic modelling for the extraction of terms related to named beaches

Juan Rojas-Garcia*, Pamela Faber**

*University of Granada
juanrojas@ugr.es
**University of Granada
pfaber@ugr.es
http://lexicon.ugr.es/faber

Abstract. EcoLexicon is a terminological knowledge base on environmental science whose design permits the geographic contextualization of data. For the geographic contextualization of landform concepts, this paper presents a semi-automatic method for extracting terms associated with named beaches (e.g., *Sumiyoshi Beach*, in Japan). Terms were extracted from an English specialized corpus on Coastal Engineering, where named beaches were automatically identified. Statistical procedures were applied for selecting terms and beaches in distributional semantic models and a topic model to construct the conceptual structures underlying the usage of named beaches. The beaches sharing associated terms were also automatically clustered and represented in the same conceptual network. The results showed that the method successfully described the semantic frames for named beaches with explanatory adequacy, according to the premises of Frame-based Terminology. Furthermore, the semantic networks unveiled the thematic relation of named beaches, coasts and rivers mentioned in the Coastal Engineering corpus to sediment transport in rivers, and the negative effects of sediment supply decrease on coastal erosion because of human activities.

1. Introduction

EcoLexicon is a multilingual, terminological knowledge base (TKB) on environmental science (http://ecolexicon.ugr.es) that is the practical application of Frame-based Terminology (Faber 2012). Since most concepts

designated by environmental terms are multidimensional (Faber 2011), the flexible design of EcoLexicon permits the contextualization of data so that they are more relevant to specific subdomains, communicative situations, and geographic areas (León-Araúz *et al*. 2013). However, the geographic contextualization of concepts related to named landforms, such as beaches (e.g., *Sumiyoshi Beach*), rivers (e.g., *Nile River*), and bays (e.g., *San Francisco Bay*) is barely tackled in terminological resources because of two reasons in our opinion : (1) they are considered mere instances of concepts such as BEACH, RIVER, or BAY, and their specific relational behaviour with other concepts in a specialized knowledge domain is thus neglected and not semantically described ; (2) their semantic representation depends on knowing which terms are related to each named landform, and how these terms are related to each other, a time-consuming task taking into account that terminologists do not often resort to natural language processing systems beyond corpus tools such as Sketch Engine (Kilgarriff *et al*. 2004).

Consequently, this paper presents a semi-automatic method of extracting terms associated with named beaches, as types of landform, from a corpus of English Coastal Engineering texts. The aim is to represent that knowledge in semantic networks in EcoLexicon according to the theoretical premises of Frame-based Terminology. Hence, on the hypothesis that named beaches should be considered concepts rather than instances in the Coastal Engineering domain, each named beach should appear in the context of a specialized semantic frame that highlights both its semantic relation to other terms and the relations between those terms.

The rest of this paper is organized as follows. Section 2 provides motivations for the research, and background on distributional semantic models and topic modelling. Section 3 explains the materials and methods applied in this study, namely, the automatic identification of named beaches, the selection procedures for terms and beaches in distributional semantic models, the clustering technique for beaches sharing associated terms, and the topic model for both the extraction of terms associated with each named beach and the construction of its corresponding specialized semantic frame. Section 4 shows the results obtained. Finally, Section 5 discusses the results, and presents the conclusions derived from this work as well as plans for future research.

2. Theoretical framework

2.1. Motivation for the research

Despite the fact that named landforms, among other named entities, are frequently found in specialized texts on environment, their representation and inclusion in knowledge resources has received little research attention, as evidenced by the lack of named landforms in terminological resources for the environment such as DiCoEnviro[1], GEMET[2] or FAO Term Portal[3]. In contrast, AGROVOC[4] basically contains a list of named landforms with hyponymic information, whereas ENVO[5] provides descriptions of the named landforms with only geographic details, and minimal semantic information consisting of the relation *located_in* (and *tributary_of* in the case of named rivers and bays).

So far, knowledge resources have limited themselves to representing concepts such as BEACH, RIVER, or BAY, on the assumption that the concepts linked to each of them are also appropriate, respectively, to all instances of named beaches, rivers, and bays in the real world. This issue is evident in the following description of forcing mechanisms acting on suspended sediment concentrations (SSC) in bays and rivers.

According to Moskalski and Torres (2012), temporal variations in the SSC of bays and rivers are the result of a variety of forcing mechanisms. River discharge is a primary controlling factor, as well as tides, meteorological forcing (i.e., wind-wave resuspension, offshore winds, storm and precipitation), and human activities. Several of these mechanisms tend to act simultaneously. Nonetheless, the specific mix of active mechanisms is different in each bay and river. For example, SSC in *San Francisco Bay* is controlled by spring-neap tidal variability, winds, freshwater runoff, and longitudinal salinity differences, whereas precipitation and river discharge are the mechanisms in *Suisun Bay*. In *Yangtze River*, SSC is controlled by tides and wind forcing, whereas river discharge, tides, circulation, and stratification are the active forcing mechanisms in *York River*.

1 http://olst.ling.umontreal.ca/cgi-bin/dicoenviro/search_enviro.cgi
2 https ://www.eionet.europa.eu/gemet/en/themes/
3 http://www.fao.org/faoterm/en/
4 http://aims.fao.org/en/agrovoc
5 http://www.environmentontology.org/Browse-EnvO

Therefore, in a knowledge resource, a set of such forcing mechanisms semantically linked to the BAY and RIVER concepts would not represent the knowledge really transmitted in specialized texts. To cope with this type of situation, terminological knowledge bases should include the semantic representation of named landforms.

To achieve that aim in EcoLexicon regarding named beaches, the knowledge should be represented in a semantic network according to the theoretical premises of Frame-based Terminology (Faber 2012), which propose knowledge representations with explanatory adequacy for enhanced knowledge acquisition in communicative situations such as specialized translation (Faber 2009). Hence, on the hypothesis that named beaches should be considered concepts rather than instances, each named beach should appear in the context of a specialized semantic frame that highlights both its relation to other terms and the semantic relations between those terms. The construction of these semantic networks and the semi-automatic extraction of terms from a specialized corpus are described in this paper. As far as we know, this framework has not been studied in the context of specialized lexicography, which is an innovative aspect of this work.

2.2. Distributional semantic models

Distributional semantic models (DSMs) represent the meaning of a term as a vector, based on its statistical co-occurrence with other terms in the corpus. According to the distributional hypothesis, semantically similar terms tend to have similar contextual distributions (Miller and Charles 1991). The semantic relatedness of two terms is estimated by calculating a similarity measure of their vectors, such as Euclidean distance, or cosine similarity, *inter alia*.

Depending on the language model (Baroni *et al*. 2014), DSMs are either count-based or prediction-based. Count-based DSMs calculate the frequency of terms within a term's context (i.e., a sentence, paragraph, document, or a sliding context window spanning a given number of terms on either side of the target term). The Correlated Occurrence Analogue to Lexical Semantics (COALS) (Rohde *et al*. 2006) is an example of this type of model.

Prediction-based models mostly exploit neural probabilistic language models, which represent terms by predicting the next term on the basis of previous terms. Examples of predictive models include the continuous bag-of-words (CBOW) and skip-gram (SG) models (Mikolov *et al*. 2013).

DSMs have been used in combination with clustering (i.e., automatic classification of objects into groups based on shared features). Work on lexical semantics applying DSMs and clustering techniques includes the identification of semantic relations (Bertels and Speelman 2014), word sense discrimination and disambiguation (Pantel and Lin 2002), automatic metaphor identification (Shutova *et al*. 2010), and classification of verbs into semantic groups (Gries and Stefanowitsch 2010).

2.3. Topic modelling for text mining

Probabilistic topic modelling is a machine learning technique that automatically identifies themes or topics in a given corpus (Blei 2012). This digital technology allows to explore documents based on the topics that run through them, rather than on keywords search alone. Latent Dirichlet Allocation (LDA) (Blei *et al*. 2003) is the approach to topic modelling that has been most frequently employed. The following explanation of topic models describes LDA and is largely based on Griffiths and Steyvers (2004), Murakami *et al*. (2017), and Spies (2018).

In topic modelling, each term in each corpus document is assigned to one topic. A document thus consists of multiple topics of different probability (e.g., 20 percent Topic A, 10 percent Topic B, 5 percent Topic C, and so forth), approximately following the proportion of terms in the document that are assigned to each topic. All documents in a corpus share the same set of topics, but with different proportions. Therefore, a document can deal with multiple topics, and the terms that appear in that document reflect the particular set of topics it addresses.

In natural language processing, the way of modelling the contributions of different topics to a document is to treat each topic as a probability distribution over terms, viewing a document as a probabilistic mixture of these topics. Starting from observed data in a corpus (i.e., occurrence frequency of the terms in the documents), LDA is able to infer a latent structure from the corpus, consisting of a set of topics.

The content of a topic is thus reflected in the terms to which it assigns high probability. For example, high probabilities for «woods», «hill» and «stream» would suggest that a topic refers to the countryside, whereas high probabilities for «check», «bank» and «credit» would suggest that a topic refers to finance.

From a cognitive view, topic modelling can be related to human capabilities to categorize documents. Psychological research found strong empirical evidence supporting cognitive adequacy of LDA, in comparison to semantic spaces such as Latent Semantic Analysis (Spies, 2018).

As DSMs, a topic model provides a form of semantic representation, a computational analogue of how human might form semantic representations through their linguistic experience. Accordingly, the association between terms can be estimated. Since the term vectors are probability distributions over topics, the relatedness is quantified by means of information-theoretic measures for probability distributions such as Hellinger distance (Csiszár and Shields 2004), or Jensen-Shannon divergence (Lee 1999), *inter alia*.

In corpus linguistics, topic models have previously been used for a variety of applications, including metaphor identification (Navarro Colorado and Tomás 2015), thematic exploration of specialized corpora (Murakami *et al*. 2017) and literary corpora (Jockers and Mimno 2013), and selectional preferences for predicate arguments (Ritter *et al*. 2010).

3. Materials and methods

3.1. Materials

3.1.1. Corpus data

The terms related to named beaches were extracted from a subcorpus of English texts on Coastal Engineering, comprising roughly 7 million tokens and composed of specialized texts (scientific articles, technical reports and PhD dissertations), and semi-specialized texts (textbooks and encyclopedias on Coastal Engineering). This subcorpus is part of the English EcoLexicon Corpus (23.1 million tokens) (see León-Araúz *et al*. (2018) for a detailed description).

3.1.2. GeoNames geographic database

The automatic detection of the named beaches and coasts in the corpus was performed with a GeoNames database dump. GeoNames (http://www.geonames.org) has over 10 million proper names for 645 different geographic entities, such as bays, beaches, coasts, rivers, lakes, and mountains. For each entity, information about their normalized designations, alternate designations,

latitude, longitude, and location name is stored. A daily GeoNames database dump is publicly available as a worldwide text file.

3.2. Methodology

3.2.1. Pre-processing

After their compilation and cleaning, the corpus texts were tokenized, tagged with parts of speech, lemmatized, and lowercased in R programming language. The multi-word terms in EcoLexicon were then automatically matched in the lemmatized corpus and joined with underscores.

3.2.2. Recognition of named beaches and coasts

Both normalized and alternate names of the beaches and coasts in GeoNames were searched in the lemmatized corpus. A total of 609 designations were recognized and listed. Nevertheless, since various designations can refer to the same beach because of morphological variation (e.g., *Egmond Beach* and *Egmond Aan Zee*), and orthographic variation (e.g., *Torrey Pines Beach* and *Torrey Pine Beach*), a procedure was created to identify variants and give them a single designation in the corpus. Because of space constraints, the procedure is not described.

Once the variants were normalized in the lemmatized corpus and joined with underscores, the number of named beaches and coasts were 294. They are shown on the map in Figure 1, with color-coded rectangles that depict their frequency in the corpus. Their latitudes and longitudes were retrieved from the GeoNames database dump. Figure 1 also reflects the representativeness of the corpus in reference to beach and coast locations.

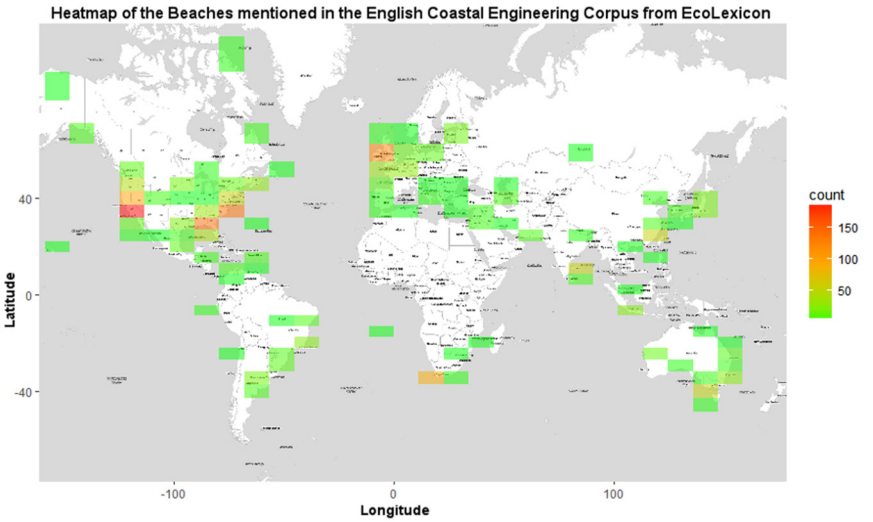


FIG. 1 – *Map with the location and color-coded frequency of the named beaches and coasts.*

The occurrence frequency of the named beaches and coasts ranged from 118 to 1 mention. In our study, only those ones with a frequency greater than 9 were considered. Figure 2 shows the 55 named beaches and coasts that fulfilled this condition, along with their numbers of mentions. Both named beaches and named coasts are henceforth referred to as named beaches for the sake of simplicity.

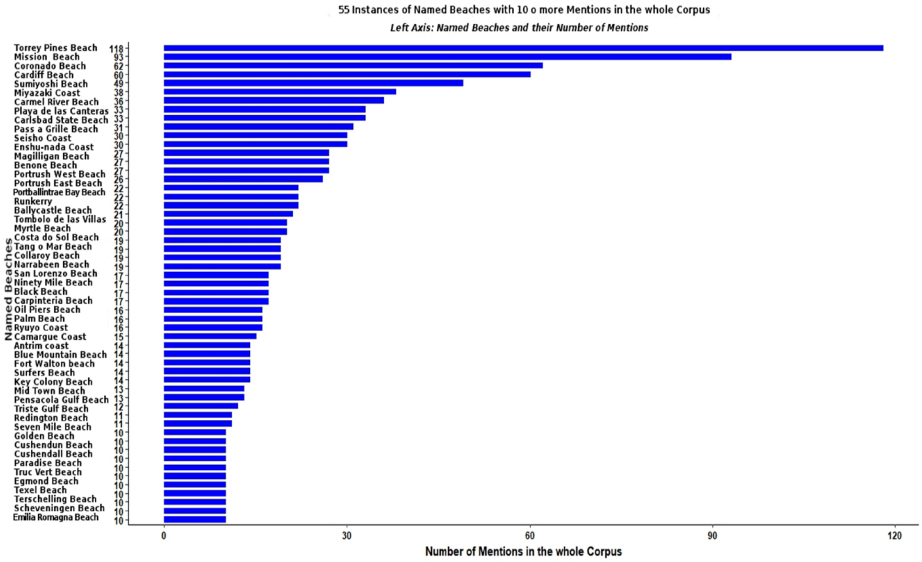


FIG. 2 – *Designations and number of mentions of the named beaches and coasts with occurrence frequency higher than 9.*

3.2.3. Term-term matrix construction

A count-based DSM was selected to obtain term vectors since this type of DSM outperforms prediction-based ones on small-sized corpora (Ars *et al.* 2016; Sahlgren and Lenci 2016).

For the construction of the DSM, only terms larger than two characters were considered. Numbers, symbols and punctuation marks were removed. Additionally, the minimal occurrence frequency was set to 5 (Evert 2009). The sliding context window spanned 30 terms on either side of the target term because large windows improve the DSM performance for small corpora (Rohde *et al.* 2006; Bullinaria and Levy 2007) and capture more semantic relations (Jurafsky and Martin 2019). We followed standard practice and did not use stopwords (i.e., determiners, conjunctions, relative adverbs, and prepositions) as context words (Kiela and Clark 2014). Since only nouns are represented in the semantic networks, adjectives, adverbs, and verbs were also disregarded as context words.

The resulting DSM was a 4431×4431 matrix *A*, whose row vectors represented the 55 named beaches plus the 4376 terms inside the context windows of 30 terms on either side of those beaches.

3.2.4. Selection of beaches and terms for clustering purposes

Subsequently, a 55×4376 submatrix *B* was extracted from *A*, where the rows represented the 55 named beaches, and the columns represented the 4376 terms co-occurring with them. To cluster the beaches in the rows of *B* sharing the same associated terms, it was necessary to select both the beaches and the terms that best discriminated different groups of beaches. This was done by removing the beaches and the terms that could act as random noise and adversely affect the clustering results (Kaufman and Rousseeuw 1990). For that aim, the method developed by Moisl (2011) was applied, succinctly explained below.

A document was considered to be the set of all context windows where a certain named beach appeared, and thus corresponded to a row of the matrix *B*. As such, we had 55 named-beach documents. Moisl's (2011) method consists in statistically ascertaining which named-beach documents are too short to estimate the probability of each term in the column vectors with a 95 % confidence level, and to remove the corresponding rows from the matrix and also the terms in the columns with low estimates.

As expected, the 29 named beaches selected by Moisl's (2011) method were those with the highest number of mentions in the corpus (Figure 2), from *Torrey Pines Beach* (118 mentions) to *Carpinteria Beach* (17 mentions). Accordingly, 310 terms were also selected. Therefore, a 29×310 submatrix *C* was extracted from *B* to group the beach vectors.

3.2.5. Clustering of named beaches and weighting schemes

The 29×310 submatrix *C* was then subjected to three weighting schemes. First, the statistical log-likelihood measure calculated the association score between all term pairs, since it captures syntagmatic and paradigmatic relations (Bernier-Colborne and Drouin 2016 ; Lapesa *et al*. 2014) and achieves better performance for small-sized corpora (Alrabia *et al*. 2014). Secondly, the scores were transformed by applying logarithms to reduce skewness (Lapesa *et al*. 2014). Finally, the row vectors were normalized to unit length.

A hierarchical clustering technique was applied to the weighted submatrix *C*. The cosine distance was used as the intervector distance measure, and the

Ward's method as the clustering algorithm (i.e., a criterion for choosing the pair of clusters to merge at each step, based on the minimum increase in total within-cluster variance).

Since it was not clear how strongly a cluster was supported by data, a means for assessing the certainty of the existence of a cluster in corpus data was devised. Multiscale bootstrap resampling (Suzuki and Shimodaira 2004) is a method for this in hierarchical clustering, which was implemented in the R package *pvclust* (Suzuki and Shimodaira 2006). For each cluster, this method produces a number ranging from zero to one. This number is the approximately unbiased probability value (AU *p*-value), which represents the possibility that the cluster is a true cluster. The greater the AU *p*-value, the greater the probability that the cluster is a true cluster supported by corpus data. An AU *p*-value equal to or greater than 95% significance level is most commonly adopted in research.

Five groups of beaches with *p*-values higher than 95% were strongly supported by corpus data, as marked by the red rectangles in the dendrogram in Figure 3.

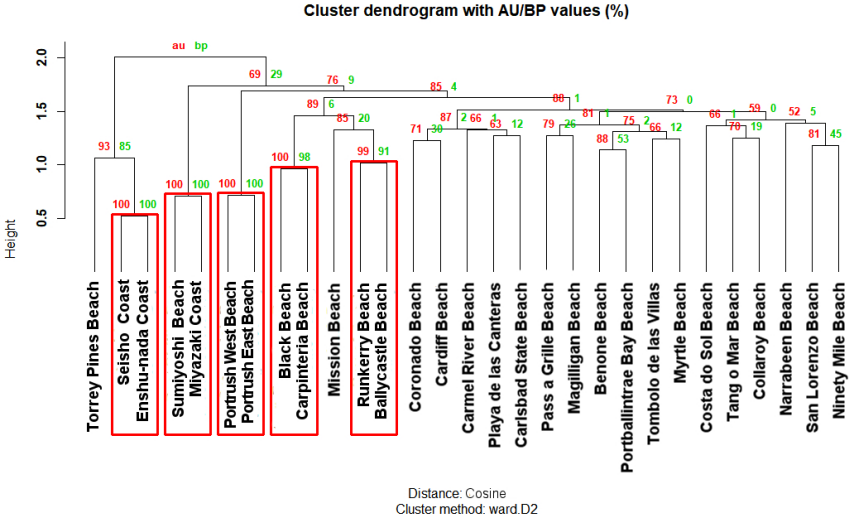


FIG. 3 – *Dendrogram of the hierarchical clustering of the 29 named beaches with 5 clusters.*

3.2.6. Selection of terms for semantic network construction

Since the 310 terms of the submatrix *C* were not sufficient to straightforwardly construct the semantic networks for the 5 clusters of beaches, another statistical method was employed to select the terms that best described the 29 beaches. In Corpus Linguistics, Moisl (2015, 77-93) suggests retaining the term columns with the highest values in four statistical criteria: *raw frequency*, *variance*, *variance-to-mean ratio* (vmr) and *term frequency-inverse document frequency* (tf-idf).

Moisl's (2015) method was applied to a 29×3838 frequency matrix, whose rows represented the 29 named beaches. The columns represented all the terms co-occurring with them (excluding the beaches). Figure 4 shows the co-plot of the four criteria, *z*-standardized for comparability reasons, and sorted in descending order of magnitude. A threshold of up to 1000 was set. This meant that only 847 terms fulfilled all criteria.

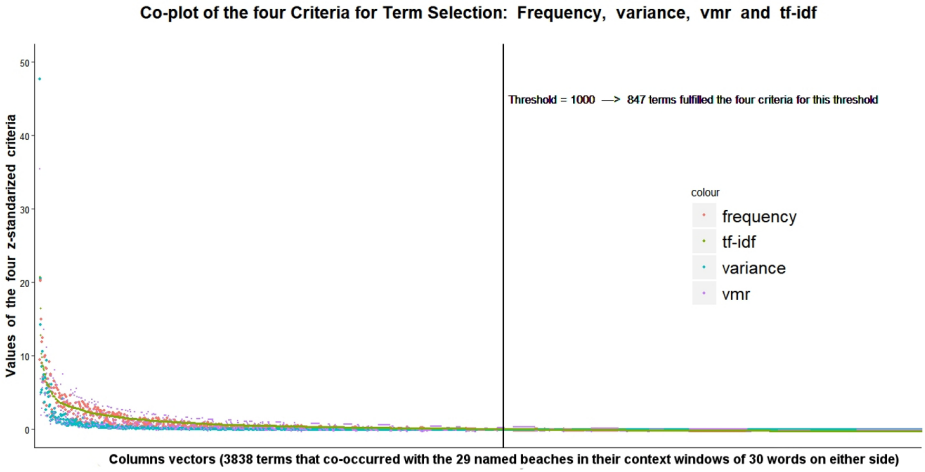


FIG. 4 – *Co-plot of the four criteria for term selection.*

3.2.7. Topic modelling for the extraction of terms associated with named beaches

Once 847 terms were selected for the semantic description of the 29 beaches, the relatedness of each beach to those terms was estimated by means of a topic model. The Biterm Topic Model (BTM) (Yan *et al.* 2013), based

on LDA, was applied to the lemmatized corpus, but containing only the occurrences of the 29 beaches and the 847 terms selected.

In our case, BTM was chosen for the following reasons: (1) it was found that BTM outperforms LDA for small corpora and short texts (note that the corpus contained only 876 term types, namely, 29 beaches plus 847 terms) because it helps to alleviate the data sparsity problem of LDA (i.e., the low co-occurrence frequency of term pairs reduces the semantic coherence of the topics) (*ibidem*); (2) BTM explicitly models the term co-occurrences in local context windows rather than in the document level, thus capturing the short-range dependencies between terms.

For BTM, a context window size of 30 terms was set, the same value as that in the DSMs used for the clustering of the beaches and the selection of the terms for the construction of semantic frames. However, the appropriate number of topics needed to be found by experimentation, calculating the harmonic mean of the document log-likelihood estimated by different models.

The harmonic mean of the document log-likelihood is a traditional measure used to select the topic model with the best generalization capability, namely, the ability of the model to identify the topics treated in unseen document, based on the analysis of the topics appearing in the documents of a training corpus. The greater the harmonic mean of the document log-likelihood, the best.

The *BTM* package for R programming language was applied to the corpus, and 39 models were computed for the topic numbers ranging from 2 to 40. The estimated harmonic mean of the document log-likelihood for each model is shown in Figure 5, where the optimal number of topics was found to be 40.

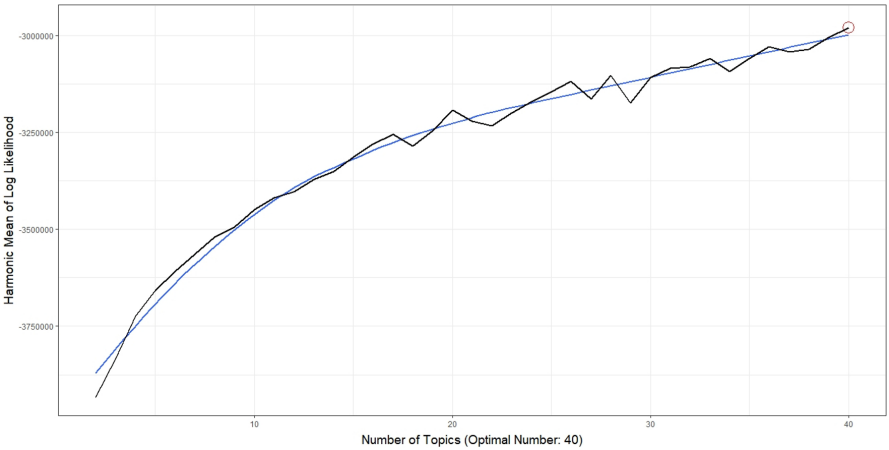


FIG. 5 – *Estimated harmonic mean of the document log-likelihood of 39 topic models, with a number of topics ranging from 2 to 40, respectively. The optimal number of topics was 40.*

An 876×40 matrix *E* was thus extracted from the topic model, where the rows represented the 847 terms plus the 29 beaches, and the columns the 40 inferred topics. Since each cell contained the probability of a term to belong to a topic, the matrix *E* is called term-topic matrix in the literature.

3.2.8. Terms characterizing each cluster

To ascertain the terms strongly associated with each of the 5 clusters, the following procedure was used:

1. For each of the named beaches in the 5 clusters, a set of the top-30 terms, most associated with each beach, was extracted from the term-topic matrix *E* using Hellinger similarity, namely, the inverse magnitude of Hellinger distance. Hellinger similarity ranges from zero to one. The greater the Hellinger similarity between two term vectors, the stronger the relatedness of the terms is considered.

2. For each cluster, the mathematical operation *set intersection* was applied to the sets of the top-30 terms, most associated with the beaches in the same cluster. Only the shared terms with a Hellinger similarity higher than 0.55 were selected.

A reduced set of terms was thus obtained for each cluster to describe the named beaches.

4. Results

Because of space constraints, only the results for the first and second clusters in Figure 3 (numbering the clusters from left to right) are presented in this paper. As shown in Figure 3, the first cluster is formed by the *Seishu* and *Enshu-nada* coasts, both located in Japan. The *Sumiyoshi Beach* and the *Miyazaki Coast*, also located in Japan, comprise the second cluster. These clusters were selected because both contain different coasts, and they are all situated in Japan. We found it interesting to explore the reasons why different coasts were grouped together, and why there were two groups of Japanese coasts in the dendrogram rather than only one.

For the description of the frames, the semantic relations were manually extracted by querying the corpus in Sketch Engine (Kilgarriff *et al*. 2004), and analyzing knowledge-rich contexts, namely, a context indicating at least one item of domain knowledge that could be useful for conceptual analysis (Meyer 2001, 281). The query results were concordances of any elements between the coast/beach in a cluster and related terms in a ±40 span. The semantic relations were those in EcoLexicon (Faber *et al*. 2009), with the addition of *supplies*, *prevents*, *accumulates_in*, *inputs*, and *simulates*.

4.1. First cluster: *Seishu* and *Enshu-nada* coasts

After the construction of dams and coastal protection structures (i.e., breakwaters, jetties, etc), and extensive *riverbed excavation* for sand mining, the sediment supplied from the *Sakawa* and *Tenryu* rivers markedly decreased, resulting in *beach erosion* on both the *Seisho* and *Enshu-nada* coasts, into which the *Sakawa* and *Tenryu* rivers discharge, respectively. Additionally, since *submarine canyons* have developed very close to the shoreline on the *Seisho Coast*, most *river sediment* from the *Sakawa River* sinks into them because of the *fluvial fan* at its mouth, thus causing *sand loss*. Since urgent measures were required to protect both coasts, beach topography changes were predicted. For that reason, the beach modifications were simulated using the *contour-line-change model* considering the following: the variation in grain size of the beach sediments, the *longshore sand transport* through the *submarine canyons*, and the sediment supply from both rivers.

In the case of the *Seisho Coast* (see Figure 6), the most favourable result was obtained when nourishment was performed using fine- and coarse-sized materials, known as *mixture materials*, because the *Seisho Coast* advanced, and the *seabed erosion* near the *submarine canyons* was prevented.

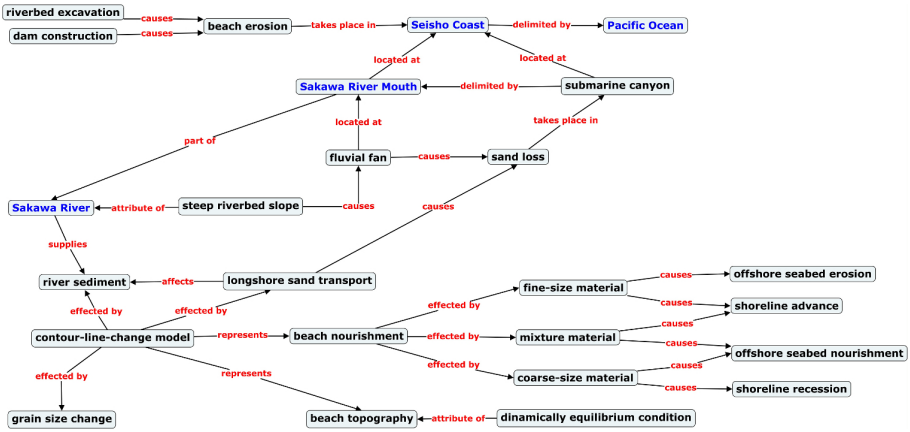


FIG. 6 – *Semantic network of the terms associated with the* Seisho Coast.

In the case of the *Enshu-nada Coast* (see Figure 7), *sand bypassing* (i.e., man-induced transfer of sand from a given distance landwards of the coastline to a beach) at *Sakuma Dam*, as a measure against *beach erosion* on the *Enshu-nada Coast*, was taken to recover the sandy beach, but *breakwaters*, previously constructed as a measure against *beach erosion*, were a barrier to the movement of sand by *longshore sand transport*.

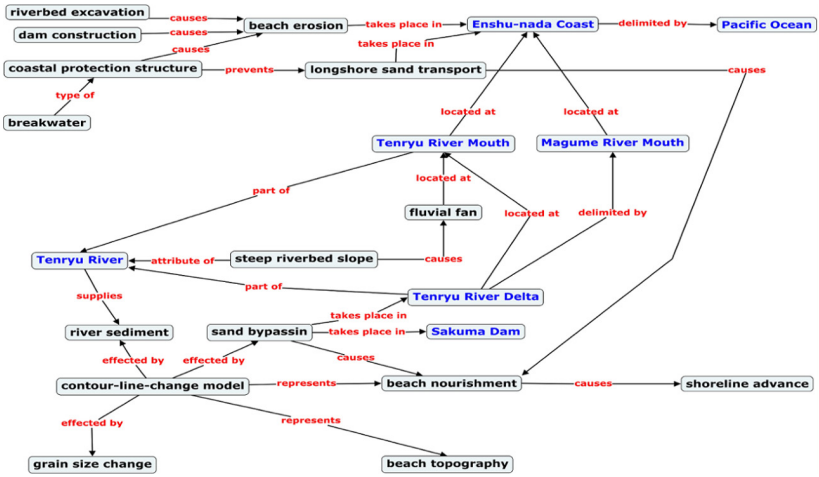


FIG. 7 – *Semantic network of the terms associated with the* Enshu-nada Coast.

4.2. Second cluster: *Sumiyoshi Beach* and *Miyazaki Coast*

Owing to the interruption of sediment flow at dams, degradation of the riverbed was observed downstream of the *Omaru*, *Mimigawa*, *Hitotsuse*, and *Ooyodo* rivers. Sediment discharge through these four rivers was thus considered to decrease considerably, causing *coastal erosion* on the *Miyazaki Coast*. The *Sumiyoshi Beach*, located on this coast, is thus a severely eroded beach because of the decrease in sediment supply from the four rivers, and the blocking of *longshore sand transport* by the *breakwater* of the *Miyazaki Port* (see Figure 8).

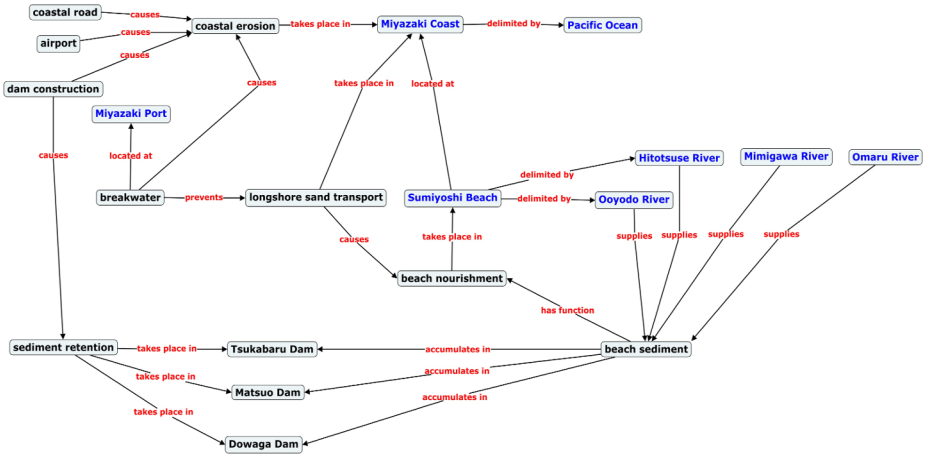


Fig. 8 – *Semantic network of the terms for the* Sumiyoshi Beach *and the* Miyazaki Coast.

5. Conclusions

To extract knowledge for the semantic frames (Faber 2012) that underlie the usage of named beaches and coasts in Coastal Engineering texts, a semi-automated method for the extraction of terms and semantic relations was devised. The semantic relations linking concepts in the semantic frames were manually extracted by querying the corpus in Sketch Engine, and analysing knowledge-rich contexts. The query results were concordances of any elements between the coast/beach in a cluster and related terms in a ±40 span. It was a time-consuming task, although essential for the explanatory adequacy of frames (Faber 2009). In future research, the automatic extraction of semantic relations for named beaches by means of knowledge patterns (KPs) (Meyer 2001) will be tested. KPs are lexico-syntactic markers that generally convey semantic relations in real texts. For instance, examples of generic-specific KPs are *such as*, *is a kind of*, and *other*. In León Araúz *et al.* (2016), a KP-based sketch grammar for Sketch Engine was developed, which automatically provides a list of terms that hold a specific semantic relation with a target term. In future work, these KPs will be applied to our corpus, as already done in Rojas-García and Cabezas-García (2019) for other purposes.

The method for the extraction of terms closely associated with named beaches offered successful results to construct semantic frames with explan-

atory adequacy, according to the premises of Frame-based Terminology. It combined, on the one hand, a count-based DSM, weighted by the log-likelihood association measure, to cluster beaches, and selection procedures for both beaches and terms based on statistical criteria. On the other hand, a topic model was employed to extract the terms related to each named beach.

The semantic networks described in the previous section reflect that most terms related to named coasts and beaches are complex nominals (e.g., *longshore sand transport*, and *beach nourishment*). English complex nominals are multi-word terms (MWTs) with a head noun preceded by a modifying element (i.e., nouns or adjectives) (Levi 1978). The abundance of MWTs is due to, at least, three reasons: (1) specialized language units are mostly represented by such compound forms (Nakov 2013); (2) complex nominals provide relevant information for the conceptual structuring of a specialized domain (Meyer and Mackintosh 1996); and (3) they are frequently used to designate specialized concepts in English (Sager *et al*. 1980). For these reasons, complex nominals should be included in the semantic networks and in TKBs such as EcoLexicon (Cabezas-García and Faber 2018).

The semantic frames also underlie the usage of named beaches and their associated terms in Coastal Engineering texts, and provide the background knowledge about them necessary in communicative situations, such as specialized translation to appropriately render terms into another language (Faber 2012). Moreover, they make the semantic and syntactic behavior of terms explicit by means of the description of semantic relations and term combinations (Faber 2009).

Finally, the conceptual structures clearly highlighted that Coastal Engineering texts attach great importance to the study of: (1) the processes that each named river triggers, (2) the processes that affect a certain named coast/beach, (3) the crucial role that a named river plays to prevent coastal erosion, and (4) the close relation between rivers and beaches in sediment transport. On the evidence of these findings supporting our working hypothesis, it would be more appropriate for named beaches, coasts, and rivers in the Coastal Engineering domain to be considered concepts for themselves rather than mere instances of the BEACH, COAST, and RIVER concepts to be semantically represented in terminological resources.

Acknowledgements

This research was carried out as part of project FFI2017-89127-P, Translation-Oriented Terminology Tools for Environmental Texts (TOTEM), funded by the Spanish Ministry of Economy and Competitiveness. Funding was also provided by the FPU grant given by the Spanish Ministry of Education to the first author. In addition, we would like to thank the anonymous reviewers for helpful discussion.

References

Ars, F., Willits, J., & Jones, M. (2016). Comparing predictive and co-occurrence based models of lexical semantics trained on child-directed speech. In A. Papafragou, D.J. Grodner, D. Mirman & J. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 1092-1097). Austin, Texas : Cognitive Science Society.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict ! A systematic comparison of context-counting vs. context-predicting semantic vectors. In K. Toutanova & H. Wu (Eds.), *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (pp. 238-247). Baltimore, Maryland : ACL, vol. 1.

Bertels, A., & Speelman, D. (2014). Clustering for semantic purposes : Exploration of semantic similarity in a technical corpus. *Terminology*, 20 (2), 279-303.

Blei, D.M. (2012). Probabilistic topic models. *Communications of the ACM*, 55 (4), 77-84.

Blei, D.M., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993-1022.

Bullinaria, J.A. & Levy, J.P. (2007). Extracting semantic representations from word co-occurrence statistics : A computational study. *Behavior Research Methods*, 39 (3), 510-526.

Cabezas-García, M., & Faber, P. (2018). Phraseology in specialized resources : An approach to complex nominals. *Lexicography*, 5 (1), 55-83.

Csiszár, I., & Shields, P.C. (2004). Information theory and statistics : A tutorial. *Foundations and Trends in Communications and Information Theory*, 1 (4), 417-528.

Evert, S. (2009). Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics. An International Handbook* (pp. 1212-1248). Berlin : Mouton de Gruyter, vol 2.

Faber, P. (2009). The cognitive shift in terminology and specialized translation. *MonTI. Monografías de Traducción e Interpretación*, 1, 107-134.

Faber, P. (2011). The Dynamics of specialized knowledge representation : Simulational reconstruction or the perception-action interface. *Terminology*, 17 (1), 9-29.

Faber, P. (Ed.) (2012). *A cognitive linguistics view of terminology and specialized language*. Berlin/Boston : De Gruyter Mouton.

Faber, P., León-Araúz, P., & Prieto, J.A. (2009). Semantic relations, dynamicity, and terminological knowledge bases. *Current Issues in Language Studies*, 1, 1-23.

Gries, S., & Stefanowitsch, A. (2010). Cluster analysis and the identification of collexeme classes. In S. Rice & J. Newman (Eds.), *Empirical and experimental methods in cognitive/functional research* (pp. 73-90). Stanford, California : CSLI.

Griffiths, T.L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101 (1), 5228-5235.

Jockers, M.L., & Mimno, D. (2013). Significant themes in 19-century literature. *Poetics*, 41 (6), 750-769.

Jurafsky, D., & Martin, J.H. (2019). Vector semantics and embeddings. In *Speech and Language Processing*. Draft of October 2, 2019. https ://web.stanford.edu/~jurafsky/slp3/6.pdf (last access : 2019-10-15).

Kaufman, L., & Rousseeuw, P. (1990). *Finding groups in data*. Hoboken, New Jersey : Wiley-Interscience.

Kiela, D., & Clark, S. (2014). A systematic study of semantic vector space model parameters. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality* (pp. 21-30). Gothenburg, Sweden : EACL.

Kilgarriff, A., Rychlý, P., Smrz, P., & Tugwell, D. (2004). The Sketch Engine. In G. Williams & S. Vessier (Eds.), *Proceeding of the 11th EURALEX International Congress* (pp. 105-115). Lorient, France : Lorient Cedex.

Lee, L. (1999). Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics* (pp. 25-32). College Park, Maryland : ACL.

León-Araúz, P., Reimerink, A., & Faber P. (2013). Multidimensional and multimodal information in EcoLexicon. In A. Przepiórkowski, M. Piasecki, K. Jassem & P. Fuglewicz (Eds.), *Computational Linguistics* (pp. 143-161). Berlin : Springer.

León-Araúz, P., San Martín, A., & Faber P. (2016). Pattern-based word sketches for the extraction of semantic relations. In P. Drouin, N. Grabar, T. Hamon, K. Kageura & K. Takeuchi (Eds.), *Proceedings of the 5th International Workshop on Computational Terminology* (pp. 73-82). Osaka : CompuTerm.

León-Araúz, P., San Martín, A., & Reimerink, A. (2018). The EcoLexicon english corpus as an open corpus in Sketch Engine. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (Eds.), *Proceedings of the XVIII EURALEX International Congress* (pp. 893-901) Ljubljana : EURALEX.

Levi, J. (1978). *The Syntax and semantics of complex nominals*. New York : Academic Press.

Meyer, I. (2001). Extracting knowledge-rich contexts for terminography : A conceptual and methodogical framework. In D. Bourigault, C. Jacquemin & M.C. L'Homme (Eds.), *Recent advances in computational terminology* (pp. 279-302). Amsterdam/Philadelphia : John Benjamins.

Meyer, I., & Mackintosh, K. (1996). Refining the terminographer's concept-analysis methods : How can phraseology help ? *Terminology*, 3, 1-26.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *Workshop Proceedings of International Conference on Learning Representations*. Scottsdale, Arizona : ICLR.

Miller, G.A., & Charles W.G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6 (1), 1-28.

Moisl, H. (2011). Finding the minimum document length for reliable clustering of multi-document natural language corpora. *Journal of Quantitative Linguistics*, 18 (1), 23-52.

Moisl, H. (2015). *Cluster analysis for corpus linguistics*. Berlin : De Gruyter Mouton.

Moskalski, S., & Torres, R. (2012). Influences of tides, weather, and discharge on suspended sediment concentration. *Continental Shelf Research*, 37, 36-45.

Murakami, A., Thompson, P., Hunston, S., & Vajn, D. (2017). 'What is this corpus about ?': Using topic modelling to explore a specialised corpus. Corpora, 12 (2), 243-277.

Nakov, P. (2013). On the interpretation of noun compounds : Syntax, semantics, and entailment. *Natural Language Engineering*, 19 (3), 291-330.

Navarro Colorado, B., & Tomás, D. (2015). A fully unsupervised topic modeling approach to metaphor identification. In *Actas del XXXI*

Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (without pagination). Alicante, Spain : SEPLN.

Pantel, P., & Lin, D. (2002). Discovering word senses from text. In *Proceedings of ACM Conference on Knowledge Discovery and Data Mining* (pp. 613-619). Edmonton, Canada : KDD-02.

Ritter, A., Mausam, & Etzioni, O. (2010). A latent Dirichlet allocation method for selectional preferences. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 424-434). Uppsala, Sweden : ACL.

Rohde, D., Gonnerman, L., & Plaut, D. (2006). An improved model of semantic similarity based on lexical co-occurrence. *Communications of the ACM*, 8, 627-633.

Rojas-García, J. & Cabezas-García, M. (2019). Use of knowledge patterns for the evaluation of semiautomatically-induced semantic clusters. In I. Simonnæs, Ø. Andersen & K. Schubert (Eds.), *New Challenges for Research on Language for Special Purposes* (pp. 121-140). Berlin : Frank & Timme.

Sager, J.C., Dungworth, D., & McDonald, P.F. (1980). *English special languages. Principles and practice in science and technology.* Wiesbaden : Brandstetter Verlag.

Sahlgren, M., & Lenci, A. (2016). The effects of data size and frequency range on distributional semantic models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 975-980). Austin, Texas : ACL.

Shutova, E., Sun, L., & Korhonen, A. (2010). Metaphor identification using verb and noun clustering. In C.R. Huang & D. Jurafsky (Eds.), *Proceedings of the 23rd International Conference on Computational Linguistics* (pp.1002-1010). Beijing, China : COLING, vol. 2.

Suzuki, R. & Shimodaira, H. (2006). Pvclust : An R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, 22 (12), 1540-1542.

Spies, M. (2018). Probabilistic topic models for small corpora – An empirical study. In C. Roche (Ed.), *TOTh 2017 – Terminologie & Ontologie : Théories et Applications* (pp. 137-160). Chambéry, France : Éditions de l'Université Savoie Mont Blanc.

Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013). A biterm topic model for short texts. In *Proceedings of the 22nd International Conference on World Wide Web* (pp. 1445-1456). Rio de Janeiro, Brazil : WWW.

Attribute-based Approach to Hyponymic Behavior in Botanical Terminology

Juan Carlos Gil-Berrozpe

University of Granada – Department of Translation and Interpreting
Buensuceso 11, 18071 Granada, Spain
jcgilberrozpe@ugr.es
http://lexicon.ugr.es/gil

Abstract. Attributes are basic to conceptualization because they expand the meaning of concept types, such as entities and events. They are often a constituent part of multi-word terms (MWTs), which represent specialized concepts in a given knowledge domain. Since attributes contain hyponymic nuances that make MWTs different from the single-word terms to which they are linked, hyponymy is intrinsically associated with the phenomenon of MWT formation. This paper presents a corpus-driven study that was performed to explore the hyponymic behavior of botanical terminology from an attribute-based approach. Additionally, a semantic analysis was carried out to distinguish the codified semantic relations and the hyponymic nuances of the attributes in botanical MWTs. Finally, based on the data, the most relevant hyponymy subtypes in the botanical corpus were assessed. Our results showed that describing the semantic information provided by attributes could lead to a more comprehensive representation of hyponymic MWTs in lexicographic and terminological resources.

1. Introduction

From an ontological point of view, attributes are one of the most basic elements of conceptualization because they expand the meaning of other concept types, such as entities and events. Attributes make concept definitions more specific by adding specific characteristics or features. Additionally, they also map out the semantic relations of a concept by associating it with other concepts and providing information about its combinatorial potential

(Faber, 2009). In this sense, and depending on their role as specifiers of concept meaning, attributes are usually nouns codifying properties or relations, or adjectives that can be qualifying or relational (Rodríguez-Pedreira, 2000).

Because they are modifiers, attributes are often a constituent of multi-word terms (MWTs), which designate specialized concepts in a given knowledge domain. One example of an MWT is *involucral bract*, where the head (*bract*) is modified and complemented by the preceding modifier (*involucral*), and which is a new concept with more specific characteristics. In this line, MWTs can be regarded as hyponyms of the single-word term (SWT) acting as a head, which is the hypernym. In other words, there is a generic-specific relation between both elements because the MWT inherits the properties of the hypernym, and further specifies its meaning by adding a new set of characteristics. Since attributes contain hyponymic nuances that differentiate MWTs from the SWTs to which they are linked, hyponymy is intrinsically associated with the phenomenon of MWT formation.

This paper presents a corpus-driven study that explores the hyponymic behavior of botanical terminology from an attribute-based approach. Additionally, a semantic analysis was carried out to distinguish the codified semantic relations and the hyponymic nuances of the attributes in botanical MWTs. Finally, based on the extracted data, the most relevant hyponymy subtypes present in the botanical corpus were assessed.

2. Attributes, multi-word terms and hyponymy : representation and description

MWTs are expressions with a head noun modified by one or more attributes, which are either nouns (i.e. an entity or a process) or adjectives (i.e. a qualifying or a relational characteristic). In the case of English, they reflect the tendency to express knowledge in semantically condensed structures (Levi, 1978 ; Štekauer *et al*., 2012 ; Sanz-Vicente, 2012). Thus, a modifier stacked on the left is known as pre-modification, which is the most frequent formation pattern in English (Cabezas-García & León-Araúz, 2019), whereas a modifier located at the right of the head is known as post-modification. An important aspect of MWTs is the fact that they present semantic patterns that can be analyzed in order to describe the semantic relation between the modifying attribute and the head (Cabezas-García & Gil-Berrozpe, 2018).

In this line, the semantic relation that is at the core of MWT formation is hyponymy, also known as the generic-specific relation. It is a unidirectional

relation in which child concepts inherit the properties of their parent concepts (Gil-Berrozpe *et al*. 2018). Moreover, this property inheritance between generic concepts and specific concepts is commonly represented through the explicitation of the *genus* (i.e. the hypernym or superordinate) and the *differentiae* (i.e. the characteristics that distinguish each hyponym) (Gheorghita & Pierrel, 2012). Therefore, the attributes of an MWT contain specific hyponymic nuances that are linked to the different subtypes of hyponymy (Gil-Berrozpe *et al*., 2017). As in the case of many MWTs, these attributes can be analyzed in terms of predicate nominalization or predicate deletion (Levi, 1978). For example, *stem photosynthesis* can be described as "photosynthesis that is performed by the stem", which reveals an agentive nuance provided by the attribute.

Accordingly, the attributive information in MWTs can be property-based or relation-based. Property-based attributes refer to inherent characteristics of concepts (e.g. the shape or color of a leaf), whereas relation-based attributes establish semantic relations with other concepts (e.g. the ecosystem in which a given species can thrive) (Gagné, 2000). Attributive information in MWTs can thus be used to identify hyponymy, as some of its subtypes are evidently attribute-based. In this respect, it is important to describe the internal semantic relations in MWTs and their implicit knowledge patterns (Meyer 2001), because they reflect how the elements of the micro-context are associated (Cabezas-García & Gil-Berrozpe, 2018).

To narrow the spectrum and highlight the most representative examples, this research focused on exploring the hyponymic nuances of frequent attributes in botany, an attribute-rich knowledge field in environmental science. Botanical terminology is characterized by a multitude of qualifying and relational adjectives that describe the various biological aspects of flora and fungi (i.e. morphology, anatomy, physiology, classification, etc.). The identification and categorization of plant species has traditionally been based on the description of visually perceptible features (Pitkänen-Heikkilä, 2015). Whilst botanical terminology has a well-established international nomenclature for flora, botanical attributes are not standardized and depend on the salient features of different types of plants. Flora is usually classified in general categories with common properties that branch out into specialized concepts with more specific characteristics. The hyponymic relations in botany thus stem from the intrinsic attributes of concrete entities and properties common to all category members.

3. Materials and methods

As a general overview, this study used a corpus-driven methodology to explore the hyponymic behavior of botanical terminology. This method was used to extract and select the most relevant hyponymic MWTs for the subsequent semantic analysis of the hyponymic nuances codified in their attributes.

For the purpose of our study, a 4-million-word English corpus containing botany-related academic literature (i.e. flora catalogs, botany manuals, annual plant reviews, plant guidebooks, specialized journals) was compiled and uploaded to the EcoLexicon English corpus in the EcoLexicon[1] terminological knowledge base (Faber *et al*., 2016 ; San-Martín *et al*., 2017). EcoLexicon is a dynamic terminological resource that targets the acquisition of environmental knowledge. This is possible thanks to an accessible and intuitive user interface that facilitates the access to conceptual, linguistic, semantic, phraseological, and multimodal information about each concept. EcoLexicon currently has approximately 4,500 environmental concepts and 23,500 terms in seven languages (English, Spanish, German, French, Dutch, Modern Greek, and Russian), with future plans to include terms in Chinese and Arabic. Its information covers various subdomains of environmental science and engineering, such as geology, oceanography, marine engineering, waste management, and botany.

The 4-million-word English botany corpus was uploaded to the EcoLexicon English corpus, and subsequently made available for corpora exploitation through Sketch Engine[2], a cloud-based corpus management and analysis tool (Kilgarriff *et al*., 2004, 2014). For example, Sketch Engine allows users to perform advanced searches within their own corpora by building expressions in Corpus Query Language or CQL (Jakubíček *et al*. 2010). Another feature of Sketch Engine are word sketches, which are automatic corpus-based summaries of a word's grammatical and collocational behavior (Kilgarriff *et al*., 2004). In addition, it is possible to create customized word sketches through the implementation of grammars containing dedicated CQL expressions (León-Araúz & San Martín, 2018).

The first step was to extract the 50 most frequent SWTs by using Sketch Engine's word list function. In this initial phase, only nouns were extracted because we wished to study the influence of their modifying attributes. The

1 https://ecolexicon.ugr.es/
2 https://www.sketchengine.eu/

frequency of the candidates ranged from 12,000 (e.g. *apex*, 12342; *stem*, 12196; *leaf*, 7299) to 400 (e.g. *phytochrome*, 481; *stipule*, 459; *bracteole*, 429), as shown in Table 1.

Term	Freq.	Term	Freq.	Term	Freq.
apex	12342	peduncle	1474	seta	853
stem	12196	anther	1465	auxin	839
leaf	7299	xylem	1348	phloem	771
lobe	4048	photosynthesis	1332	cation	771
bract	3886	spore	1322	cannabinoid	755
almond	3060	ligule	1305	rhizome	718
achene	3039	meristem	1263	rootstock	702
inflorescence	2862	filament	1111	gibberellin	669
subspecies	2774	corolla	1045	nutlet	641
lamina	2658	hilum	1036	whorl	638
capitulum	2392	receptacle	1024	peristome	611
petiole	2390	angiosperm	974	sporophyte	594
stamen	1990	pappus	956	umbel	534
calyx	1986	pedicel	941	phytochrome	481
cultivar	1865	chloroplast	900	stipule	459
costa	1741	sepal	863	bracteole	429
midrib	1512	tuft	853		

Tab. 1 – *Extraction of the 50 most frequent SWTs (exclusively nouns) in the botany corpus.*

However, since our goal was to analyze hyponymic MWTs, it was necessary to find the most relevant MWTs for each SWT. For this reason, the following CQL query was performed for each of the 50 SWTs extracted:

[tag="N.*|JJ.*|RB.*|VVG.* |VVN.*"]{1,}[lemma="apex"]

The element on the right of the query is the nominal head of the MWT. In this case, it is specified as [lemma="apex"] because it refers to the SWT *apex*. The element between quotation marks was replaced by each of the previous SWTs in order to perform each query. This nominal head can be preceded by nouns (N.*), adjectives (JJ.*), adverbs (RB.*), present participles (VVG.*) or past participles (VVN.*) on the order of one or more modifiers ({1,}). This CQL expression was based on the elements that modify nouns in English MWTs, since pre-modification is the preferred formation pattern in this lan-

guage (Cabezas-García & León-Araúz, 2019). After applying this query, the results in Sketch Engine were filtered by node form frequency (Figure 1).

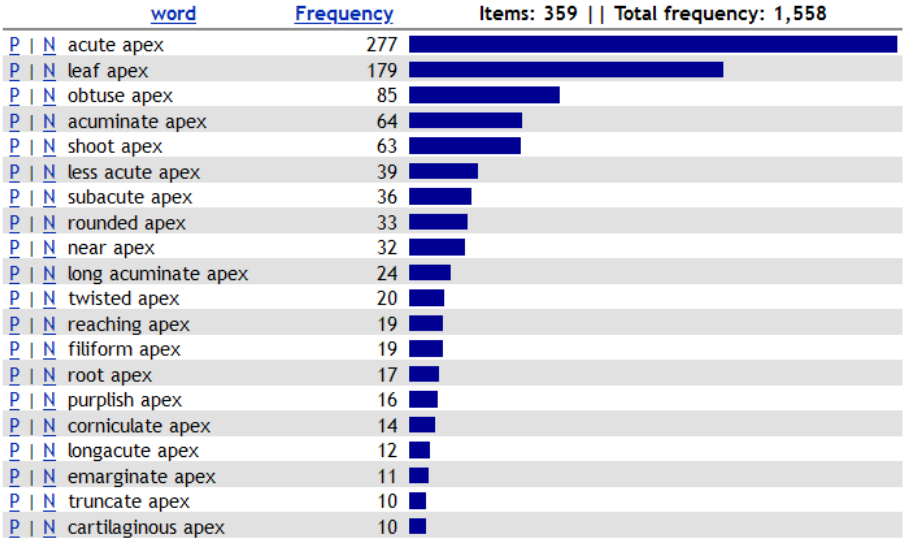


FIG. 1 – *Extraction of the 20 most frequent MWTs related to the SWT* apex *and filtered by node form frequency.*

The 20 most frequent MWTs of each query were thus extracted, because each case presented different frequencies and values for the CQL query performed. In total, we extracted 500 hyponymic MWTs associated with the 50 hypernymic SWTs. Table 2 shows a sample which contains three examples (*stem*, *rootstock*, and *xylem*) of the final MWT extraction process.

Hypernym	Freq.	Term	Freq.	Term	Freq.
stem	**12196**	**rootstock**	**702**	**xylem**	**1348**
Hyponyms	Freq.	Hyponyms	Freq.	Hyponyms	Freq.
flowering stem	123	almond rootstock	48	secondary xylem	37
main stem	106	peach rootstock	30	developing xylem	25
secondary stem	70	seedling rootstock	21	primary xylem	10
upper stem	60	citrus rootstock	19	stem xylem	10
primary stem	57	hybrid rootstock	18	embolized xylem	9
photosynthetic stem	38	prunus rootstock	13	root xylem	9
fertile stem	36	stout rootstock	12	functional xylem	7
woody stem	35	clonal rootstock	10	conifer xylem	5
sterile stem	32	resistant rootstock	10	leaf xylem	5
branched stem	19	dwarfing rootstock	9	new xylem	5
young stem	19	plum rootstock	8	shoot xylem	4
new stem	15	fruit rootstock	8	old xylem	4
shrub stem	14	new rootstock	8		
erect stem	13	lime rootstock	8		
tree stem	13	branched root-stock	6		
plant stem	13	woody rootstock	6		
aerial stem	12	tough rootstock	5		
lower stem	12	oblique rootstock	5		
underground stem	12	commercial root-stock	5		
leafy stem	9	tuberous rootstock	4		

Tab. 2 – *Sample of the extraction of the 500 most relevant hyponymic MWTs.*

4. Results and discussion

The 500 hyponymic MWTs were semantically analyzed based on the following parameters: (1) the SWT acting as hypernym; (2) the MWT acting as hyponym; (3) the frequency of the MWT; (4) number of modifiers in the MWT; (5) bracketing of the MWT to identify its core element when there was more than one modifier; (6) the type of attribute (property-based or relation-based); (7) the semantic relation in the MWT; (8) the hyponymic nuance contained in the modifying attribute of the MWT; and (9) the hyponymy subtype expressed by the MWT. Because of length restrictions, this paper

focuses on the four most representative case studies. Afterwards, the overall results and findings of the analysis are shown.

4.1. Semantic analysis of hyponymic MWTs

4.1.1. Case study 1 : *stem photosynthesis*

Table 3 shows the results for the MWT *stem photosynthesis*, a hyponym of *photosynthesis*. Accordingly, *stem photosynthesis* is the synthetization process of organic compounds from carbon dioxide and water using light energy, which is carried out by the chlorophyll of the stem. Therefore, the attribute is relation-based because the representation of this MWT can be simplified as "the photosynthesis that is PERFORMED BY the stem". Because of the semantic relation codified in this MWT, the attribute *stem* contains an agentive hyponymic nuance by which *stem photosynthesis* is related to *photosynthesis* through agent-based hyponymy.

(1) Hypernym (SWT)	photosynthesis
(2) Hyponym (MWT)	stem photosynthesis
(3) Frequency of the MWT	78
(4) Number of modifiers	1
(5) Bracketing	-
(6) Type of attribute	relation-based
(7) Codified semantic relation	PERFORMED _ BY
(8) Hyponymic nuance in the attribute	agentive
(9) Hyponymy subtype	agent-based hyponymy

TAB. 3 – *Semantic analysis of the botanical MWT* stem photosynthesis.

4.1.2. Case study 2 : *outer involucral bract*

Table 4 shows the results for *outer involucral bract*, hyponym of *bract*. In this case, it was necessary to bracket the MWT to distinguish the attribute (*outer*) from the head (*involucral bract*). An *outer involucral bract* is the external group of whorls beneath a flower or a flower cluster. Thus, this attribute is also relation-based because its expression is codified as "the involucral bract that is LOCATED AT the outside part". Because of this implicit semantic relation, the attribute *outer* contains a locative hyponymic nuance by which *outer involucral bract* is a subtype of *bract* based on its location.

(1) Hypernym (SWT)	bract
(2) Hyponym (MWT)	outer involucral bract
(3) Frequency of the MWT	554
(4) Number of modifiers	2
(5) Bracketing	outer [involucral bract]
(6) Type of attribute	relation-based
(7) Codified semantic relation	LOCATED _ AT
(8) Hyponymic nuance in the attribute	locative
(9) Hyponymy subtype	location-based hyponymy

TAB. 4 – *Semantic analysis of the botanical MWT* outer involucral bract.

4.1.3. Case study 3 : *divergent branch leaf*

Table 5 shows the case study of *divergent branch leaf*, a hyponym of *leaf*. Bracketing was performed to divide the attribute (*divergent branch*) and the head (*leaf*) of the MWT. In this line, a *divergent branch leaf* is the flat green blade attached to the stem of a plant whose shape branches in different directions. Thus, in this case the attribute is property-based because it is referring to an intrinsic and differentiating characteristic of the hyponym. Its representation can be simplified as "a leaf that HAS A divergent branch SHAPE". Due to the semantic relation codified in this MWT, the attribute *divergent branch* contains a hyponymic nuance that expresses shape and by which *divergent branch leaf* is associated with *leaf* through a shape-based hyponymy.

(1) Hypernym (SWT)	leaf
(2) Hyponym (MWT)	divergent branch leaf
(3) Frequency of the MWT	77
(4) Number of modifiers	2
(5) Bracketing	[divergent branch] leaf
(6) Type of attribute	property-based
(7) Codified semantic relation	HAS _ ATTRIBUTE _ SHAPE
(8) Hyponymic nuance in the attribute	shape
(9) Hyponymy subtype	shape-based hyponymy

TAB. 5 – *Semantic analysis of the botanical MWT* divergent branch leaf.

4.1.4. Case study 4 : *flowering stem*

Table 6 shows the MWT *flowering stem*, hyponym of *stem*. In this respect, a *flowering stem* is defined as the type of stem that is capable of forming flowers. As in case study 3, the attribute in the MWT is property-based because it expresses a unique feature of the hyponym that makes it differ from other types of *stem*. Therefore, its meaning can be semantically described as "a stem that HAS A flowering ABILITY". Because of the semantic relation in this MWT, *flowering* has a hyponymic nuance that describes an ability, and *flowering stem* is thus a subtype of *stem* according to an ability-based hyponymy.

(1) Hypernym (SWT)	stem
(2) Hyponym (MWT)	flowering stem
(3) Frequency of the MWT	123
(4) Number of modifiers	1
(5) Bracketing	-
(6) Type of attribute	property-based
(7) Codified semantic relation	HAS _ ATTRIBUTE _ ABILITY
(8) Hyponymic nuance in the attribute	ability
(9) Hyponymy subtype	ability-based hyponymy

TAB. 6 – *Semantic analysis of the botanical MWT* flowering stem.

4.2. Overall results

After carrying out the semantic analysis of the 500 botanical MWTs, it was possible to obtain valuable insights into the behavior of the attributes and their influence on hyponymy. In relation to the nature of the attributes, 287 (57 %) were property-based and 213 (43 %) were relation-based (see Figure 2). This indicates that there is a general trend in botanical terminology to prefer property-based attributes or, in other words, those that refer to the inherent characteristics of a concept. This is hardly surprising, since botany is a field in which the physical traits related to the morphology and anatomy of flora and fungi are very important. However, quite interestingly, relation-based attributes were present in 43 % of the cases analyzed, which reveals a recurrent set of semantic relations in botanical MWTs.

Fig. 2 – *Property-based and relation-based attributes of the 500 botanical MWTs*

As for the hyponymic nuances of the botanical MWTs, Figure 3 displays the 19 nuances that were identified: ability, agent, color, combination, component, composition, content, gender, layer, length, location, method, origin, quantity, shape, size, taste, texture, and time. These hyponymic nuances were associated with their respective hyponymy subtypes and, accordingly, the most salient subtypes were shape-based hyponymy (95, 19%), component-based hyponymy (83, 16.6%), location-based hyponymy (59, 11.8%), and color-based hyponymy (57, 11.4%). On the one hand, two of these hyponymy subtypes relied on property-based attributes, including cases based on shape nuances (e.g. *ovoid nutlet*, *branched sporophyte*, *spathe-like leaf*) and on color nuances (e.g. *golden ligule*, *reddish costa*, *pink anther*). On the other hand, the other two hyponymy subtypes were determined by relation-based attributes, with examples associated with components (e.g. *leaf apex*, *axillary peduncle*, *dorsal stamen*) and location (e.g. *upper stamen*, *central lobe*, *outer pedicel*).

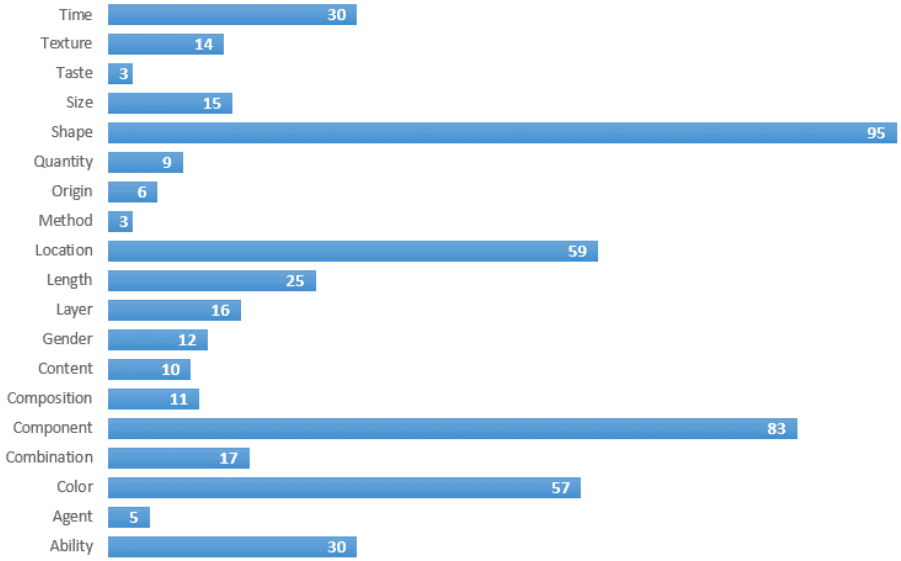


FIG. 3 – *Hyponymic nuances and hyponymy subtypes expressed by the 500 botanical MWTs*

5. Conclusion

This research combined corpus-driven methodology and semantic analysis to analyze the hyponymic behavior of botanical terminology from an attribute-based perspective, focusing on hyponymic nuances and their related hyponymy subtypes. It was found that describing the semantic information provided by attributes could lead to a more comprehensive representation of hyponymic MWTs in lexicographic and terminological resources.

The results of this study indicated that both property-based and relation-based attributes are common in botany, and that they express hyponymic nuances ranging from physical features of the head of the MWT (e.g. shape, color, size) to interactions between the head of the MWT and its modifiers (e.g. location, agents, components). However, there is a general tendency in botanical terminology to rely on property-based attributes rather than on relation-based attributes, which is the main reason why the majority of hyponymy

subtypes are related to properties. Accordingly, these are the attributes that tend to codify visual and natural characteristics of concepts, which are the cornerstone of studies on the anatomy, morphology and physiology of flora and fungi.

Our future work will focus on the description of the semantic information codified in hyponymic MWTs in EcoLexicon. In addition, further studies will enhance the description and representation of hyponymic MWTs. These will include an ontological analysis of how the conceptual categorization of botanical MWTs influences their representation through hyponymy, as well as a syntactic analysis of hyponymic knowledge patterns that most commonly express generic-specific relations in botanical terminology.

Acknowledgements

This research was carried out as part of project FFI2017-89127-P, *Translation-Oriented Terminology Tools for Environmental Texts* (TOTEM), funded by the Spanish Ministry of Economy and Competitiveness. Funding was also provided by the FPU grant (ref. FPU16/02194) given by the Spanish Ministry of Education and Professional Training to the author of the article. Additional funding was provided by a research grant (*Beca de Iniciación a la Investigación – Plan Propio de Investigación 2017*) given by the Vice-Rectorate for Research and Knowledge Transfer of the University of Granada to the author of the article.

References

Cabezas-García, Melania and Gil-Berrozpe, Juan Carlos. 2018. "Semantic-based Retrieval of Complex Nominals in Terminographic Resources". In *Proceedings of the XVIII EURALEX International Congress : Lexicography in Global Contexts*, 269-281. Ljubljana : Ljubljana University Press.

Cabezas-García, Melania and León-Araúz, Pilar. 2019. "On the Structural Disambiguation of Multi-word Terms". In *Computational and Corpus-Based Phraseology*, edited by Gloria Corpas Pastor and Ruslan Mitkov, *Lecture Notes in Computer Science*, 11755 : 46-60. Cham : Springer.

Faber, Pamela. 2009. "The Cognitive Shift in Terminology and Specialized Translation." *MonTI (Monografías de Traducción e Interpretación)*, 1 : 107-134. Valencia : Universitat de València.

Faber, Pamela ; León-Araúz, Pilar ; and Reimerink, Arianne. 2016. "EcoLexicon : New Features and Challenges". In *Proceedings of*

GLOBALEX 2016 : Lexicographic Resources for Human Language Technology in conjunction with the 10th edition of the Language Resources and Evaluation Conference, 73-80. Portorož : ELRA.

Gagné, Christina L. 2000. "Relation-Based Combinations Versus Property-Based Combinations : A Test of the CARIN Theory and the Dual-Process Theory of Conceptual Combination". *Journal of Memory and Language*, 42, 365-389.

Gheorghita, Inga and Jean-Marie**,** Pierrel. 2012. **"**Towards a methodology for automatic identification of hypernyms in the definitions of large-scale dictionary**"**. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, 2614-2618**.** Istanbul : ELRA.

Gil-Berrozpe, Juan Carlos ; León-Araúz, Pilar ; and Faber, Pamela. 2018. "Subtypes of Hyponymy in the Environmental Domain : Entities and Processes". In *TOTh 2016 (Terminologie & Ontologie : Théories et Applications)*, *Terminologica*, edited by Christophe Roche, 39-54. Chambéry : Éditions de l'Université Savoie Mont Blanc.

Gil-Berrozpe, Juan Carlos ; León-Araúz, Pilar ; and Faber, Pamela. 2017. "Specifying Hyponymy Subtypes and Knowledge Patterns : A Corpus-based Study". In *Proceedings of eLex 2017 Conference : Electronic Lexicography in the 21st Century*, 63-92. Brno : Lexical Computing CZ s.r.o.

Jakubíček, Miloš ; Kilgarriff, Adam ; McCarthy, Diana ; and Rychlý, Pavel. 2010. "Fast Syntactic Searching in Very Large Corpora for Many Languages". In *Proceedings of PACLIC 2010 : 24th Pacific Asia Conference on Language, Information and Computation*, 741-747. Sendai : Tohoku University.

Kilgarriff, Adam ; Baisa, Vít ; Bušta, Jan ; Jakubíček, Miloš ; Kovář, Vojtěch ; Michelfeit, Jan ; Rychlý, Pavel ; and Suchomel, Vít. 2014. "The Sketch Engine : ten years on". *Lexicography* 1 : 7-36.

Kilgarriff, Adam ; Rychlý, Pavel ; Smrz, Pavel ; and Tugwell, David. 2004. "The Sketch Engine". In *Proceedings of the Eleventh EURALEX International Congress*, edited by Geoffrey Williams & Sandra Vessier, 105-115. Lorient : Université de Bretagne-Sud.

León-Araúz, Pilar and San Martín, Antonio. 2018. "The EcoLexicon Semantic Sketch Grammar : from Knowledge Patterns to Word Sketches". In *Proceedings of the LREC 2018 Workshop Globalex 2018 : Lexicography & WordNets*, 94-99. Miyazaki : Globalex.

Levi, Judith. 1978. *The Syntax and Semantics of Complex Nominals*. New York : Academic Press.

Meyer, Ingrid. 2001. “Extracting Knowledge-rich Contexts for Terminography : A Conceptual and Methodological Framework”. In *Recent Advances in Computational Terminology*, edited by Didier Bourigault, Christian Jacquemin and Marie-Claude L’Homme, 279-302. Amsterdam/Philadelphia : John Benjamins.

Pitkännen-Heikkilä, Kaarina. 2015. “Adjectives as terms”. *Terminology*, 21(1), 76-101. Amsterdam/Philadelphia : John Benjamins.

Rodríguez-Pedreira, Nuria. 2000. “Adjectifs qualificatifs et adjectifs relationnels : étude sémantique et approche pragmatique”. PhD diss., Universidad de Santiago de Compostela.

San Martín, Antonio ; Cabezas-García, Melania ; Buendía, Míriam ; Sánchez-Cárdenas, Beatriz ; León-Araúz, Pilar ; and Faber, Pamela. 2017. “Recent Advances in EcoLexicon”. *Dictionaries : Journal of the Dictionary Society of North America* 38(1): 96-115.

Sanz-Vicente, Lara. 2012. “Approaching Secondary Term Formation Through the Analysis of Multiword Units : An English–Spanish Contrastive Study. *Terminology*, 18(1): 105-127. Amsterdam/Philadelphia : John Benjamins.

Štekauer, Pavol ; Valera, Salvador ; and Körtvélyessy, Lívia. 2012. *Word-Formation in the World’s Languages*. Cambridge/New York : Cambridge University Press.

Résumé

Les attributs sont essentiels à la conceptualisation puisqu’ils élargissent le sens des types de concepts, comme les entités et les événements. Ils font souvent partie intégrante des noms composés (NC), qui représentent les concepts spécialisés dans des domaines de connaissance. Étant donné que les attributs contiennent des nuances hyponymiques qui rendent les NC différents des noms simples auxquels ils sont liés, l’hyponymie est intrinsèquement associée au phénomène de formation des NC. Cet article présente une étude de corpus qui a été réalisée pour explorer le comportement hyponymique de la terminologie botanique à partir d’une approche basée sur les attributs. En outre, une analyse sémantique a été effectuée pour distinguer les relations sémantiques codifiées et les nuances hyponymiques des attributs des NC botaniques. Enfin, à partir des données, les sous-types d’hyponymie les plus importants du corpus botanique ont été évalués. Les résultats obtenus ont montré que la description de l’information sémantique fournie par les attributs pourrait conduire à une représentation plus complète des NC hyponymiques dans les ressources lexicographiques et terminologiques.

TermFrame : Knowledge frames in Karstology

Katarina Vrtovec, Špela Vintar, Amanda Saksida, Uroš Stepišnik

Faculty of Arts, University of Ljubljana, Aškerčeva 2, SI – 1000 Ljubljana
E-mail : {katarina.vrtovec, spela.vintar, amanda.saksida,
uros.stepisnik} @ff.uni-lj.si

Abstract. We describe the TermFrame knowledge base for Karstology developed in accordance with the frame-based approach in Terminology for three languages, English, Slovene and Croatian. After describing the annotation framework, annotation process and validation, we interpret the results relying on the connection between the relation frames and four most frequent semantic categories in our corpora. The goal of the research was to establish ideal frames for the aforementioned semantic categories : on the basis of the results of the research and the "ideal" set of relations for each semantic category, which were developed by a domain expert and a terminologist prior to the research, we created definition templates for our major semantic categories which we exemplify on selected karst terms. In the last part, we present a visual representation for the selected terms and propose a method to complement our domain representation with information extraction from entire corpora, thus examining all knowledge-rich contexts, not just definitions.

1. Introduction

In light of numerous studies in the field of terminology in the past decades we have come to the conclusion that traditional specialised dictionaries fail to accommodate all aspects of concepts as abstract units of knowledge embedded into communicative settings. Terminology science today understands knowledge as represented in texts as conceptually dynamic and linguistically varied (Cabré, 1999), and the cognitive frames underlying specialised communication being context-, language- and culture-dependant (Faber & Medina-Rull, 2017).

Termbases in digital environment which extend the concept-oriented approach can accommodate different cognitive layers of terminological entries, linking concepts presented as nodes, and generic and domain-specific relations between them. The frame-based approach in terminology (Faber *et al*., 2005 ; Faber, 2009 ; Faber *et al*., 2012) introduces a new framework for representing specialised knowledge : the units of knowledge, which have a multidimensional nature, are presented in a specific conceptual organisation, and all semantic and syntactic features are corpus-based (Faber, 2009). A well-known implementation of the aforementioned approach is the EcoLexicon[1], a multilingual knowledge base developed for Environmental Science. EcoLexicon in based on the environmental event which is characterised by a template with a set of prototypical conceptual relations.

The TermFrame project was inspired by the EcoLexicon research group and the frame-based approach in terminology. However, the existing methodologies were adapted and extended in order to (Vintar *et al*., 2019):

- Build a comprehensive structured knowledge base for the domain of Karstology in three languages, English, Slovene and Croatian ;
- Develop modes of knowledge representation which can be used by linguists, terminologists, experts and data scientists alike, and which adequately show language- and context-dependent differences between knowledge frames ;
- Explore new methods of knowledge extraction from specialized texts, so that our results can be generalized and applied to new languages and domains.

Our interdisciplinary team consists of a domain expert, terminologists, a cognitive linguist and a group of researchers in knowledge discovery and data mining.

At the time of writing this article, several studies within the TermFrame project are underway for all three languages. However, certain tools, such as the definition extractor, have yet to be adapted for Croatian. The results of the research in this paper are therefore for English and Slovene only.

This paper is structured as follows : Section 2 describes the resources we used, including tools for term and definition extraction, and semantic annotation. Section 3 further describes the semantic annotation used in TermFrame project, while in Section 4 we outline the frames for English and Slovene

1 http://ecolexicon.ugr.es/en/index.htm

definienda on the basis of the analysed data. Section 5 focuses on "ideal" definition frames for karst terms in English and Slovene. We conclude with the modelling of domain representations and its visualisation in section 6.

2. The TermFrame corpus and definition extraction

For the purposes of the research, we built English, Slovene and Croatian specialised corpora. All three corpora are comprised of relevant contemporary works on Karstology and are comparable in terms of domain and text types included. The corpora include scientific texts (scientific papers, books, articles, doctoral and master's theses, glossaries and dictionaries) from the field of the interdisciplinary scientific domain of Karstology encompassing geomorphology, geology, hydrology, speleology, biology etc.

	English	Slovene	Croatian
Tokens	2,386,075	1,208,240	1,229,368
Words	1,968,509	987,801	969,735
Sentences	87,713	51,990	53,017
Documents	54	60	43

TAB. 1 : TermFrame corpora

For definition extraction we used the Clowdflows definition extractor (Pollak *et al*. 2012). The definition candidates were extracted automatically with a pattern-based setup for English and Slovene. The definition candidates were later manually validated and only examples with valuable explanatory information about karst concepts were retained (yield ~ 20 %). All definitions types (intensional, extensional, functional, paraphrase etc.) were considered, therefore not all obtained definitions have the traditional structure : the definiendum may appear in different positions in the sentence, the genus may or may not be present, the term may be defined only though its hyponyms etc. After validation the yield was 215 and 259 definitions for English and Slovene respectively.

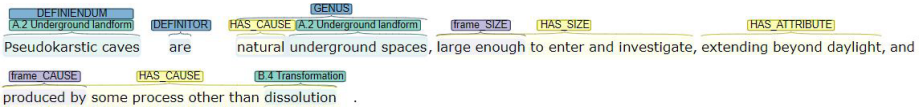


FIG. 1 : Intentional (genus-differentia definition type) definition

FIG. 2 : Extensional definition

3. The annotation framework

The annotation framework was developed prior to the beginning of our research by a terminologist and a domain expert. The framework is based on the concept of karstological event (similar to EcoLexicon's Environmental Event): landforms and other tangible object in karst represent *patients*, natural or human elements stand as *agents* and environmental influences as *processes*. The challenges of the framework are most importantly to accurately represent knowledge structures in the Karstology domain, to model visual representation of knowledge adapted for experts, terminologists and lay users respectively, and to provide training data for text mining and knowledge extraction, which would ultimately serve for the creation of the TermFrame knowledge base.

The annotation framework consists of five layers :

1. *Definition element* outlines the crucial information of every definition : DEFINIENDUM (the terms which is being defined), DEFINITOR (the defining phrase of the definition, usually a verbal phrase), GENUS (the hypernym or superordinate term), and SPECIES (the hyponym or subordinate term ; relevant in extensional definitions).
2. *Semantic category* is a taxonomically organized conceptual hierarchy adapted to Karstology (on the basis of EcoLexicon conceptual hierarchy) with the five top-level nodes (See FIG. 3).

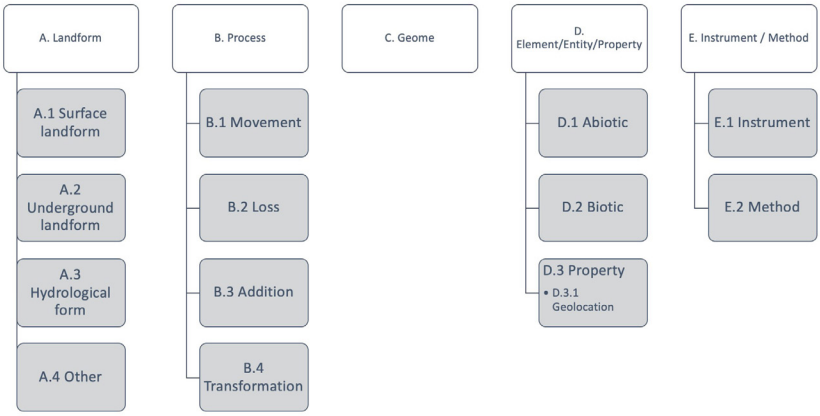


FIG. 3 : Semantic categories

The concepts represented by the categories were modelled according to the basic karstologic approach (Ford and Williams, 2007 ; Jennings, 1985) corresponding to surface and underground karst landforms (Landform) and a number of related processes (Process). Other categories included larger karst environments (Geome), materials, processes and landforms from general language, but closely connected to karst environments (Entity/Element/Property) and typical methods and tools in Karstology (Instrument/Method).

3. *Relations* mark the part of definition where a specific property of the definiendum is described. We used a set of 16 relations which may span over several words or phrases, but do not necessarily overlap with the two previous layers. The following relations were defined by domain experts according to the geomorphologic analytical approach (Pavlopoulos *et al*., 2009) considering spatial distribution (HAS_LOCATION ; HAS_POSITION), morphography (HAS_FORM ; CONTAINS), morphometry (HAS_SIZE), morphostructure (OCCURS_IN_MEDIUM ; COMPOSED_OF), morphogenesis (HAS_CAUSE), morphodynamics (AFFECTS ; HAS_RESULT ; HAS_FUNCTION), and morphochronology (OCCURS_IN_TIME). Additional relations were applied for general

properties (HAS_ATTRIBUTE; DEFINED_AS), and for research methods (STUDIES; MEASURES).

4. *Relation_definitor* may be present within a certain relation and is annotated in order to facilitate future knowledge extraction experiments. It marks the word or phrase which introduces a specific type of relation (eg. HAS_SIZE [from 1 to several metres in diameter], frame_SIZE [in diameter]).
5. *Term_canonical* was added primarily for term normalization purposes in elliptic constructions, e. g. in *water discharge and velocity* we may add *water velocity* as the canonical or full version of the term.

The semantic annotation was performed in WebAnno, an open source server-based tool which allows users to specify the annotation layers, attributes and tagsets and perform annotation, curation and monitoring (De Castilho *et al*., 2014). Each definition was annotated by two persons (linguists) and later curated by a domain expert. Regular meetings of annotators and curators were organised to discuss and evaluate the annotation procedure.

4. Frames for English and Slovene definienda

In our research, we analysed the four most common categories of karst definienda: (i) surface landforms, (ii) underground landforms, (iii) hydrological landforms and (iv) geomes (see FIG. 4 and FIG. 5).

Row Labels	Count of Category
A. Landform	1
A.1 Surface landform	73
A.2 Underground landform	45
A.3 Hydrological form	12
B. Process	2
B.1 Movement	1
B.2 Loss	1
B.3 Addition	1
B.4 Transformation	3
C. Geome	39
D.1 Abiotic	17
D.2 Biotic	3
D.3 Property	9
E.2 Method	8
Grand Total	**215**

FIG. 4: Semantic categories of definienda in English corpus

Row Labels	Count of Category
A.1 Surface landform	119
A.2 Underground landform	36
A.3 Hydrological form	20
A.4 Other	1
B. Process	6
B.1 Movement	2
B.2 Loss	1
B.4 Transformation	7
C. Geome	63
D.1 Abiotic	15
D.3 Property	11
D.3.1 Geolocation	1
E.1 Instrument	2
E.2 Method	2
Grand Total	**286**

FIG. 5 : Semantic categories of definienda in Slovene corpus

In order to model specific knowledge, we analysed the relations that co-occur with most frequent categories in our dataset. We learned that the definienda of all semantic categories are most commonly defined through location in both languages – HAS_LOCATION is the most frequent relation, followed by HAS_FORM outlining the relations describing the shape of the defined term, and HAS_CAUSE specifying the information on the origin of the defined term (See FIG. 6 and FIG. 7).

Hydrological landforms are often connected with HAS_FUNCTION in both languages – this relation is almost undetectable with other semantic categories, especially in English definitions. As far as landforms are concerned, hydrological landforms are the ones with the most unusual set of relations : they are predominantly defined though two relations, HAS_LOCATION and HAS_FUNCTION.

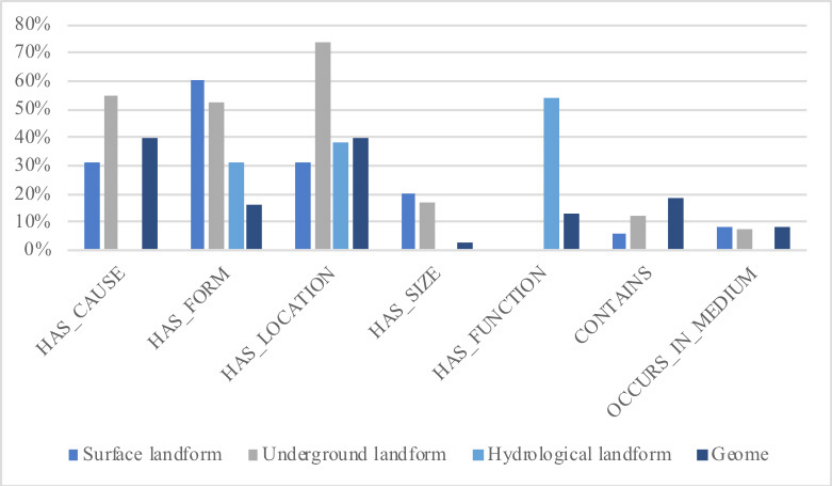


FIG. 6: Frames in English definienda

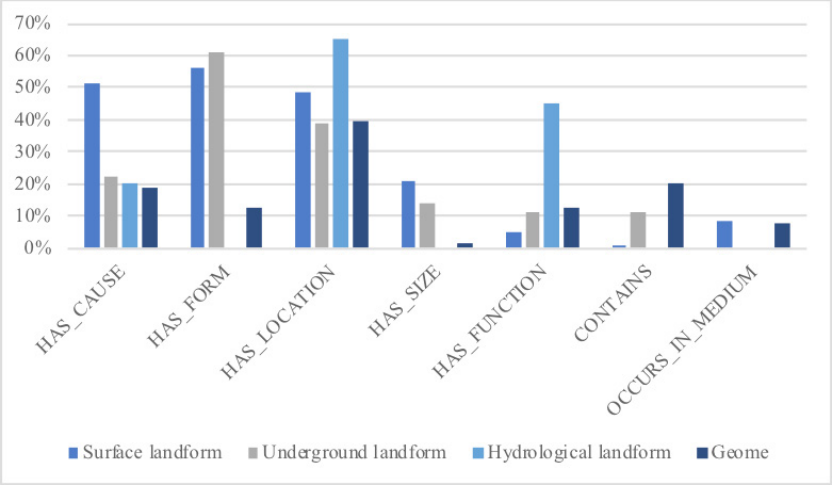


FIG. 7: Frames in Slovene definienda

The set of relations appearing with geomes is somewhat different from other semantic categories. Apart from the HAS_LOCATION and HAS_CAUSE relations, geomes are most commonly defined by the relation

CONTAINS indicating a different pattern of definition structure. Geomes being larger geographical entities (such as karst, epikarst, phreatic zone etc.) are often defined through their constituent parts (see FIG. 8).

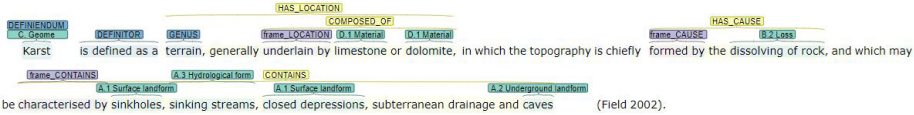


FIG. 8 : Definition for geome "karst"

5. "Ideal" definition templates for karst terms in English and Slovene

Several previous studies (Faber Benitez *et al*. 2005, Duran-Muñoz 2016) show that certain types of concepts are typically defined with specific frames. Such "ideal" frames or definition templates have been used by terminographers or experts for definition writing. In our case, karstologists seem to use explicit definition templates for some categories such as geomorphological karst features or landforms, but not for others. Our intention is to identify the typical frames reflecting cognitive structures in each language for the main concept categories in karstology, compare them between languages and possibly use them for our target knowledge base.

In the initial phase of the TermFrame project, we proposed the ideal frames for four semantic categories : (i) karst landforms, (ii) karst processes, (iii) geomes and (iv) instruments and methods. Here we focus only on karst landforms and geomes :

- A [karst landform] is a [GENUS] which HAS_FORM, HAS_SIZE, HAS_CAUSE, OCCURS_IN_MEDIUM, HAS_FUNCTION, OCCURS_IN_TIME ;
- A [geome] is a [GENUS] which CONTAINS, is COMPOSED_OF, HAS_LOCATION.

In the analysis of the definienda, our aim was to establish a set of relations specific for each semantic category ultimately helping us to predict "the frame" or "behaviour" of karst terms in authentic contexts. The results show that the frames are very similar in both languages with some variation in frequency, thus implying that the domain structure might be language – or culture – spe-

cific to a certain extent, but conceptualisation within specialised language depends on both language specific and cognitive factors. We can also draw parallels between the results of the analysis and the established ideal frames, especially for geomes (HAS_LOCATION and CONTAINS being amongst the most frequent relations) and for surface landforms (HAS_FORM, HAS_CAUSE, HAS_SIZE, OCCURS_IN_MEDIUM with the exception of HAS_LOCATION being the ideal frame according to the pattern we designed prior to the research). As expected, the results of the research include a larger spectrum of relations as we wanted to create the broadest possible background for each semantic category. However, the concepts in both languages do not fully overlap, e. g. the terms we found in English texts tend to emphasize its morphodynamic aspects (karst as a process), while terms found in Slovene texts tend to focus on mophostructural or morphographic aspect (karst as a landscape).

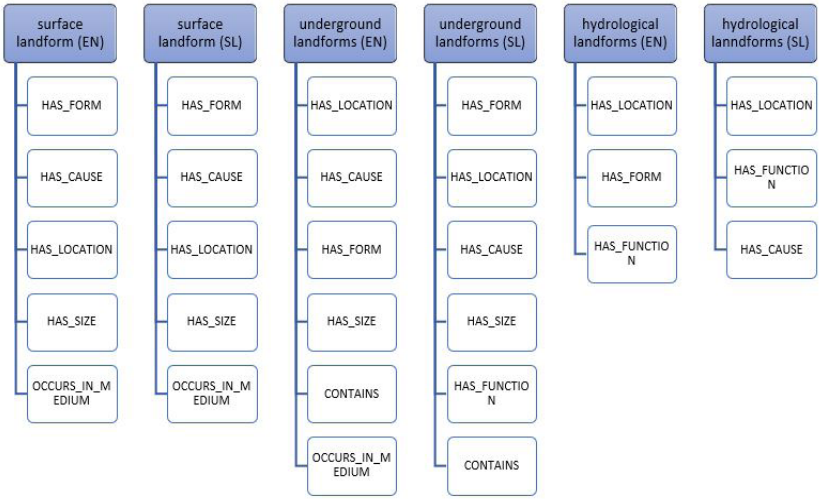


FIG. 9: Ideal frame for surface landforms, underground landforms and hydrological landforms (decreasing in frequency top-down)

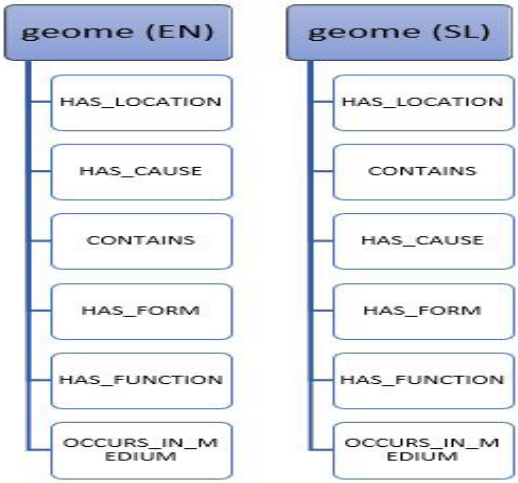


FIG. 10 : Ideal frame for geomes (decreasing in frequency top-down)

6. Domain specific knowledge modelling and visualisation

Representing domain specific knowledge in graphs allows us to interpret information faster. The concepts can be presented as nodes, which are linked to the relations between them. Users have the option to explore nodes and its neighbours, and graphs can have a multi-layered structure so that the user chooses which information they want to focus on as the representation provides an overview of the selected data.

6.1. Visualisation of data on selected terms

For plenty of key terms in karst we found several definitions with varying components, since different authors emphasise different aspects of the defined term depending heavily on the context, text type, genre etc. Our goal is to create a visual representation of the domain in order to facilitate the understanding of the relations between all knowledge elements we obtained. For illustration purposes we present graphs for two karst terms with several definitions (four or more), concentrating on the relations contained in those definitions. The first example is *sinkhole* (surface landform) and the second

karst (geome). Using the information on semantic categories and relations obtained from annotated definitions, we constructed a visual representations of knowledge with the help of the NetViz terminology[2] visualisation tool adapted for Karstology and the TermFrame project.

For *sinkhole* (see Figure 11) we found six relations in four definitions: HAS_LOCATION, OCCURS_IN_MEDIUM, HAS_FORM, HAS_ATTRIBUTE, CONTAINS and HAS_FUNCTION.

1. According to this concept, a sinkhole is part of a long-term process and it may have different forms and surface expressions at different times.
2. A typical sinkhole is bowl shaped, with one or more low spots along its bottom.
3. In order for a sinkhole to form in the first place, there are three requirements: a drainage path for the surface water runoff to follow; a zone of bedrock modified by solution located at or near the surface; and a covering of soil or some other material making up the land surface (this last is not an absolute requirement; when that cover is absent, certain types of sinkholes can still form).
4. Sinkholes or dolines are closed land surface depressions with internal drainage typically formed in karst environments (Ford and Williams, 2007).

Three of these relations comply with the "ideal" definition template for landforms (OCCURS_IN_MEDIUM, HAS_FORM, HAS_LOCATION and HAS_FUNCTION), while others (HAS_ATTRIBUTE, CONTAINS) do not. Moreover, the definitions did not contain all relations specific for landforms (HAS_SIZE, HAS_CAUSE and OCCURS_IN_TIME).

2 https://biomine.ijs.si/netviz.

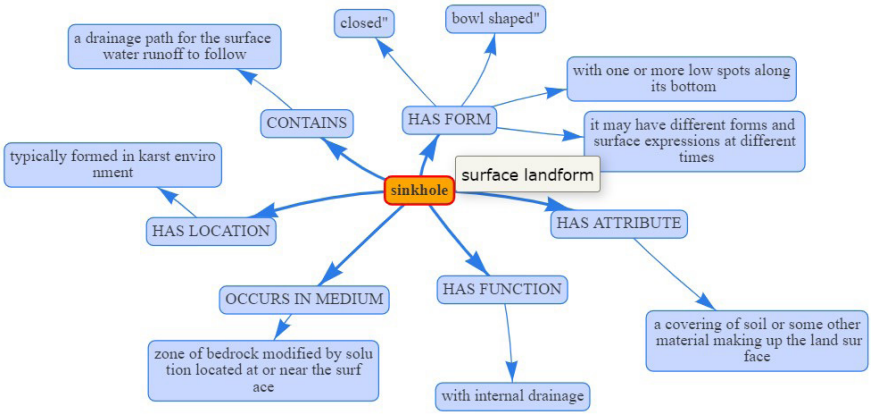


FIG. 11 : Visual representation of "sinkhole" in NetViz

For the term *karst* (see Figure 12) we were able to extract relations in perfect compliance with the ideal frame for geomes we modelled in the initial stage of the project (CONTAINS, COMPOSED_OF, HAS_LOCATION) in seven definitions.

1. Karst is a product of processes that operate on continental land masses, especially when the land masses are uplifted above sea level.
2. The term karst describes a distinctive topography that indicates dissolution (also called chemical solution) of underlying soluble rocks by surface water or ground water.
3. The term karst applies to a distinctive type of landscape that develops from the dissolving action of water on soluble bedrock, primarily limestone and marble but also dolostone, gypsum and halite.
4. Karst is defined as a terrain, generally underlain by limestone or dolomite, in which the topography is chiefly formed by the dissolving of rock, and which may be characterised by sinkholes, sinking streams, closed depressions, subterranean drainage and caves (Field 2002).

5. Karst is the term applied to terrains underlain by soluble rock: mainly limestone, dolomite and gypsum.
6. Karst is terrain with distinctive hydrology and landforms arising from the combination of high rock solubility and well-developed solution channel (secondary) porosity underground.
7. Karst is the term used to describe a special style of landscape containing caves and extensive underground water systems that is developed on especially soluble rocks such as limestone, marble, and gypsum.

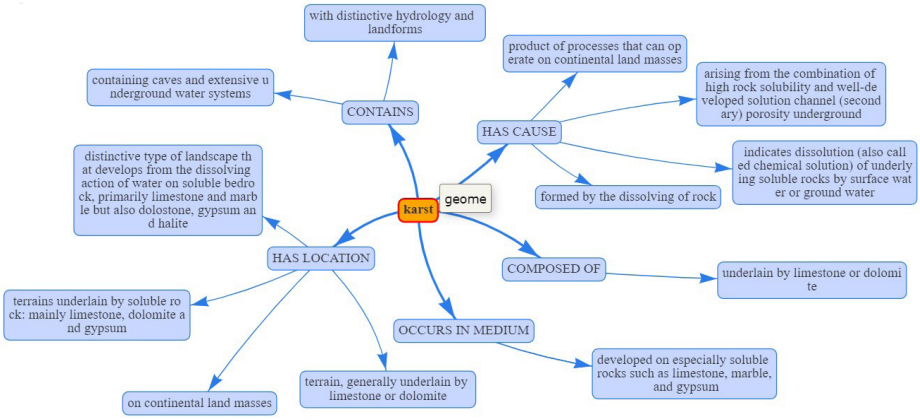


FIG. 12: Visual representation of "karst" in NetViz

The analysis of the definienda proposes additional relations for the definition templates for geomes: HAS_CAUSE and OCCURS_IN_MEDIUM, which we found in definitions for *karst*, and HAS_FORM and HAS_FUNCTION which never appeared in the definitions for this term. This may be the consequence of the selection of the term – definitions for more specific concepts in karst terminology included information on the form and function of the feature, while those for generic concepts (such as *karst*) did not.

6.2. Expanding the knowledge graph with corpus data – examining knowledge-rich contexts

Our annotation was limited to a set of definitions we extracted with the possibility of a lot of data being excluded during automatic extraction of definitions due to low recall. Furthermore, important features of karst concepts might be described in non-defining contexts.

As we were not able to complete all expected features of the ideal frame for *sinkhole* from annotated definitions alone, we complemented our research with the raw information from corpora in order to broaden the scope of the search to all knowledge-rich contexts.

Relation frames help us identify the missing pieces, e. g. the relation definitors *formed by* or *resulting from* introduce the relation HAS_CAUSE, *developed in* introduces OCCURS_IN_MEDIUM, *in diameter* introduces HAS_SIZE etc. These can prove to be helpful for the improvement of relation extraction and relation prediction tools.

Figure 13 shows the term *sinkhole* with all the features for the landform relation set.

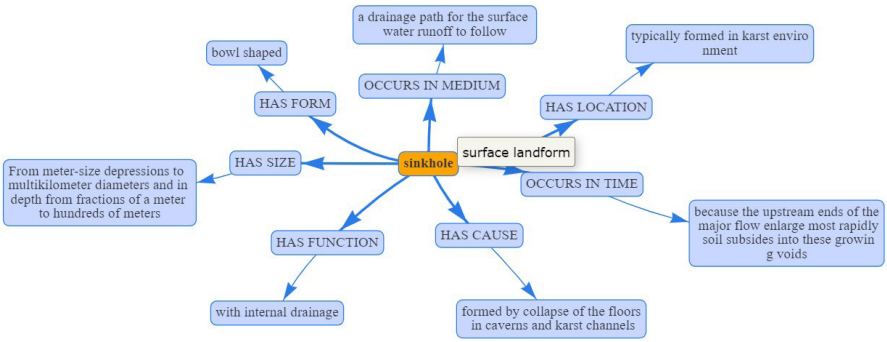


FIG. 13 : Visual respresentation for "sinkhole" complying with the ideal frame in NetViz

Visualization experiments on the basis of community detection algorithms, co-occurances and embedding-based topic modelling are underway, see Miljković *et al*. (2019).

7. Conclusion

In this paper we describe the initial results of the TermFrame approach to domain modelling. We performed a multi-layered semantic annotation on the definitions extracted from Slovene and English corpora and analysed the results based on the connection between four most frequent semantic categories and relations specifying different features of karst terms. The results show that the overall most frequent relation describes the location of the entity, followed by the relations for defining the form and the origin of the karst term. Two of the most interesting findings are that (i) karst hydrological landforms are most commonly described thought their function (whereas the relation HAS_FUNCTION rarely appears anywhere else), and (ii) geomes being larger geographical entities are frequently defined with the relation CONTAINS, which specifies a hyponymy. So far the results seem very similar for both languages, with some variation in terms of cognitive structures : English definitions tend to emphasize morphodynamic aspects of the terms, while Slovene definitions tend to focus on mophostructural or morphographic aspects.

In the last part, we structured visual domain representation for two karst terms, *sinkhole* and *karst*, and extract the missing features of the definition templates from knowledge-rich contexts from the corpus.

Our future plans go hand in hand with the notion that annotations can be used for machine learning – we plan to further explore the value of relation definitors to improve relation extraction and relation prediction tools, and to examine not only definitions, but all data from the corpora.

References

Cabré, M. T. (1999). Terminology : Theory, methods, applications. Amsterdam/ Philadelphia : J. Benjamins Publishing Company.

De Castilho, R. E., Biemann, C., Gurevych, I. and Yimam, S.M. (2014). WebAnno : a flexible, web-based annotation tool for CLARIN. In Proceedings of the CLARIN Annual Conference (CAC) 2014, Soesterberg, Netherlands.

Duran-Muñoz, I. (2016). Producing frame-based definitions. Terminology, 22/2, pp. 223-249.

Faber Benítez, P., Márquez Linares, C., & Vega Expósito, M. (2005). Framing Terminology : A process-oriented approach. Meta : Journal des traducteurs/Meta : Translators' Journal, 50(4).

Faber, P. (2009). The Cognitive Shift in Terminology and Specialized Translation. MonTI. Monografías de Traducción e Interpretación. 1 : pp. 107-134.

Faber, P., Ed. (2012). A Cognitive Linguistics View of Terminology and Specialized Language. Berlin, Boston : De Gruyter Mouton.

Faber, P., León-Araúz, P., & Reimerink, A. (2016). EcoLexicon : new features and challenges. GLOBALEX, pp. 73-80.

Ford, D. & Williams, P.D. (2007). Karst Hydrogeology and Geomorphology. Wiley, Chichester.

Jennings, J.N. (1985). Karst Geomorphology. Basil Blackwell, Oxford, pp. 293ff.

Madsen, B. N., & Thomsen, H. E. (2008). Terminological Principles Used for Ontologies. Managing Ontologies and Lexical Resources, pp. 107-122.

Miljković, D., Kralj, J., Stepišnik, U. & Pollak, S. (2019). Communities of related terms in a Karst terminology co-occurrence network. In Proceedings of eLex 2019, pp. 357-373.

Pavlopoulos, K., Evelpidou, N. & Vassilopoulos, A. (2009). Mapping Geomorphological Environments. Springer, Berlin Heidelberg.

Pollak, S., Vavpetič, A., Kranjc, J., Lavrač, N. & Vintar, Š. (2012). NLP workflow for on-line definition extraction from English and Slovene text corpora. In Proceedings of KONVENS, pp. 53--60.

Pollak, S., Repar, A., Martinc, M. & Podpečan, V. (2019). Karst exploration : extracting terms and definitions from karst domain corpus. In Proceedings of eLex 2019, pp. 934-956.

Roche, C., Calberg-Challot, M., Damas, L., & Rouard, P. (2009). Ontoterminology : A new paradigm for terminology. In International Conference on Knowledge Engineering and Ontology Development, pp. 321-326.

San Martin, A. & L'Homme, M.-C. (2014). Definition Patterns for Predicative Terms in Specialized Lexical Resources. In Proceedings of LREC14, pp. 3748-3755.

Svensén, B. (1993). Practical Lexicography : Principles and Methods of Dictionary-Making. Oxford University Press.

Temmerman, R., & Van Campenhoudt, M. (Eds.). (2014). Dynamics and Terminology : An interdisciplinary perspective on monolingual and multilingual culture-bound communication (Vol. 16). Amsterdam : John Benjamins.

Vintar, Š. (2010). Bilingual term recognition revisited : The bag-of-equivalents term alignment approach and its evaluation. Terminology, 16(2), pp. 141-158.
Vintar, Š., Saksida, A., Vrtovec, K. & Stepišnik, U. (2019). Modelling specialized knowledge with conceptual frames : the TermFrame approach to a structured visual domain representation. In Proceedings of e-Lex 2019, pp. 305-318.
Vintar, Š. & Grčić Simeunović, L. (2017). Definition frames as language-dependent models of knowledge transfer. Fachsprache 1-2/2017, pp. 43-58.
White, W.B., (1988). Geomorphology and hydrology of karst terrains. Oxford university press, Oxford.

Résumé

Nous décrivons la base de connaissances TermFrame développée selon l'approche de la terminologie des cadres pour le domaine de karstologie en trois langues : anglais, slovène et croate. Après avoir décrit le cadre et le processus d'annotation et de validation, nous interprétons les résultats en nous appuyant sur le lien entre les cadres de relation et les quatre catégories sémantiques les plus fréquentes dans nos corpus. Le but de la recherche était d'établir des cadres idéaux pour les catégories sémantiques susmentionnées. Sur la base des résultats de la recherche et de l'ensemble « idéal » de relations qui ont été développées préalablement pour chaque catégorie sémantique par un expert du domaine et un terminologue, nous proposons des modèles de définitions idéales pour des termes karstiques choisis appartenant à différentes catégories sémantiques. Dans la dernière partie, nous présentons une représentation visuelle des termes sélectionnés et proposons une méthode pour compléter notre représentation du domaine par l'extraction d'informations à partir de corpus entiers, examinant ainsi tous les contextes riches en connaissances, et non seulement les définitions.

La construction d'un domaine en perspective diachronique. Les fibres textiles chimiques aux XIX^{e} et XX^{e} siècles

Klara Dankova[1]

Università Cattolica del Sacro Cuore
Largo Gemelli 1, Milan
klara.dankova@unicatt.it

Résumé. Pour définir les termes de spécialité de manière claire et efficace, il est nécessaire de savoir quelle place occupent les concepts désignés dans le réseau structuré de relations, constituant l'organisation conceptuelle du domaine. L'objectif de cet article est de reconstruire l'évolution du domaine des fibres textiles chimiques depuis ses débuts à la fin du XIX^{e} siècle jusqu'à la fin du XX^{e} siècle, suivant l'analyse de la terminologie employée. Les fibres chimiques sont d'abord définies et situées dans la filière de l'industrie textile. L'évolution conceptuelle du domaine des fibres chimiques au sein des fibres textiles est ensuite retracée et illustrée par les changements de sens de deux termes-clés du domaine : *fibre* et *artificiel*. Enfin, la construction d'une ressource terminologique relative aux termes désignant les fibres chimiques est décrite et le mode de recensement des termes est exemplifié.

1. Introduction

Les fibres chimiques sont des matières textiles fabriquées par un processus de transformation chimique (Weidmann 2010, 14). Développées à la fin du XIX^{e} siècle comme substituts des fibres naturelles, les fibres chimiques sont désormais devenues des matières premières incontournables dans plusieurs secteurs industriels, parmi lesquels le médical, le bâtiment et l'agriculture. En raison de leur typologie très diversifiée et de leur importance croissante dans la vie quotidienne, la construction d'une ressource terminologique fiable

1 Cette étude fait partie de la thèse de doctorat sur la terminologie des fibres chimiques que je suis en train d'élaborer (dir. Maria Teresa Zanola).

s'avère être un premier pas vers une démarche de clarté définitoire dans ce domaine. Une étude approfondie du domaine de spécialité représente ainsi une condition préalable à la bonne réussite du projet.

Dans la première partie de l'article, nous allons définir la place des fibres chimiques au sein de l'industrie textile et nous allons ensuite reconstruire l'évolution conceptuelle du domaine étudié depuis sa naissance à la fin du XIXe siècle jusqu'à l'époque actuelle. L'évolution de l'organisation conceptuelle sera illustrée aussi par l'évolution sémantique de deux termes-clés du secteur : *fibre* et *artificiel*. Dans la deuxième partie, nous allons présenter les objectifs et les enjeux principaux de la construction d'une base de données des termes désignant les fibres chimiques grâce à des exemples tirés de notre corpus.

2. Organisation conceptuelle du domaine

Cette partie de l'article porte sur la place des fibres chimiques dans la production textile et sur l'évolution conceptuelle des fibres chimiques à l'intérieur du domaine des fibres textiles. Les relations entre les concepts sont représentées dans les cartes conceptuelles, que nous avons construites à l'aide du logiciel *CmapTools*.

2.1. Organisation de l'industrie textile et les fibres textiles chimiques

Les fibres textiles chimiques occupent une place importante à l'intérieur du système de production textile incluant toutes les activités qui visent à approvisionner le marché en produits textiles. Ces activités concernent essentiellement trois phases consécutives (fig. 1) :

- l'obtention de matières premières ;
- la production de structures textiles intermédiaires ;
- la fabrication de produits textiles finis.

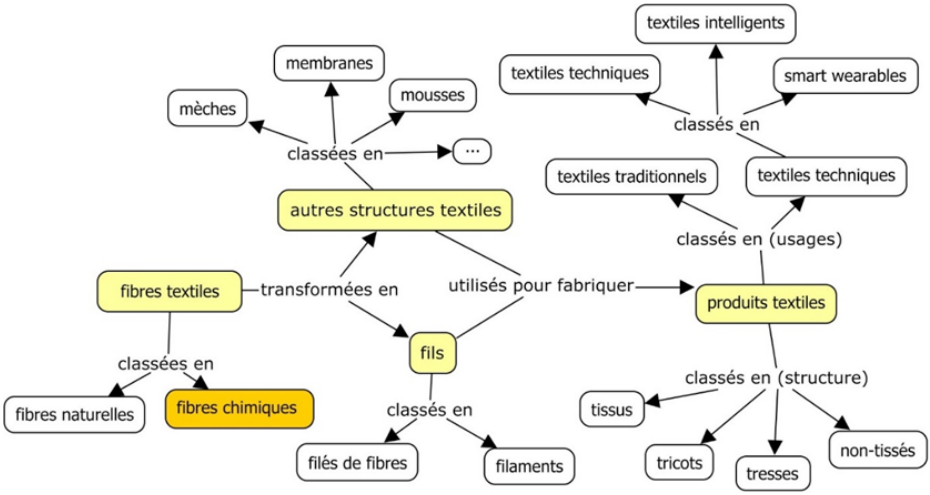


Fig. 1 – *La filière de production textile.*

Les matières premières sont appelées *fibres textiles*. Actuellement, deux catégories principales de fibres textiles sont distinguées en tenant compte de leur origine : les *fibres naturelles*, qui sont fournies directement par la nature, et les *fibres chimiques*, obtenues artificiellement par l'intervention de l'homme. Les fibres chimiques représentent donc un type de matières premières utilisées dans l'industrie textile.

Les fibres textiles servent à la production de structures textiles intermédiaires, telles que les fils, les mèches, les membranes et les mousses (Weidmann 2010), les plus importantes étant les fils. Ceux-ci sont constitués d'un *filé de fibres* ou d'un ou de plusieurs *filaments* (Weidmann 2010, 283-284). Les filés de fibres sont obtenus par la filature des fibres discontinues de courte longueur (Weidmann 2010, 284), tandis que les filaments sont fabriqués par le procédé de filage, lors duquel une matière fondue est extrudée à travers une filière (Weidmann 2010, 17). Les fibres chimiques sont fabriquées soit sous forme de fibres discontinues, soit sous forme de filaments (Agulhon 1962, 6).

Les structures textiles intermédiaires sont utilisées dans la fabrication de produits textiles finis. En fonction du mode d'assemblage des fils, les produits textiles sont classés en quatre catégories : les *tissus*, les *tricots*, les *tresses* et

les *non-tissés*. Du point de vue de leurs applications, on distingue les *textiles traditionnels*, qui sont destinés à la confection de vêtements et à l'ameublement, et les *textiles techniques* ou *textiles à usages techniques* (*TUT*), définis comme « tout produit ou matériau textile dont les performances techniques et les propriétés fonctionnelles prévalent sur les caractéristiques esthétiques ou décoratives » (Weidmann 2010, 4). Les TUT trouvent leur emploi dans de nombreux domaines d'activité, tels que le bâtiment, la défense, le sport, l'agriculture et la santé. Depuis les années 2000, deux types innovants de textiles techniques se sont développés : les *textiles intelligents*, capables de réagir aux changements de l'environnement, et les *smart wearables*, contenant des dispositifs technologiques connectés (Union des Industries Textiles 2017, 5-7).

2.2. L'évolution conceptuelle du domaine des fibres textiles

Le domaine des fibres chimiques s'est constitué à l'intérieur des fibres textiles à la fin du XIX^e^ siècle. C'est pourquoi nous allons décrire l'évolution de son organisation conceptuelle dans le cadre des fibres textiles en distinguant quatre phases principales :

- depuis l'époque préhistorique jusqu'en 1884 : l'emploi des fibres naturelles ;
- depuis 1884 jusqu'à la fin des années 1930 : l'introduction des fibres artificielles ;
- depuis la fin des années 1930 jusqu'à la fin des années 1960 : l'introduction des fibres synthétiques et des nouvelles fibres inorganiques ;
- depuis les années 1970 jusqu'à présent : l'introduction des fibres spéciales.[2]

Les phases indiquées ci-dessus ont été proposées en tenant compte de l'apparition de nouveaux types de fibres textiles, à savoir : les fibres artificielles, les fibres synthétiques et les fibres spéciales. Il faut mettre en évidence que la diffusion de fibres innovantes est relativement lente : en général, il y a une différence considérable entre la mise au point d'une fibre textile dans les laboratoires et son lancement sur le marché, voire sa production à l'échelle industrielle. À titre d'exemple, les fibres artificielles ont été développées à la fin du XIX^e^ siècle, mais elles ne deviennent des biens de consommation courante

2 La catégorie des fibres spéciales a été créée par l'auteur pour regrouper trois types de fibres, désignées par D. Weidmann (2010) sous les termes de *fibres fonctionnalisées*, *fibres spécifiques* et *fibres hybrides*.

que dans les années 1920 et 1930. De plus, il faut considérer que la constitution d'une nouvelle catégorie de fibres est souvent influencée par les stratégies de marketing appliquées par les entreprises, qui exploitent souvent le manque de clarté autour de la composition et des propriétés des produits existant sur le marché (Bonetti *et al*. 2012, 5-6). Les fibres ignifuges, par exemple, sont perçues comme une catégorie particulière environ depuis les années 1970, même si les premiers traitements ignifuges ont été brevetés déjà au cours des années 1920 (Fauque et Bramel 1999, 123).

2.2.1. Depuis l'époque préhistorique jusqu'en 1884 : l'emploi des fibres naturelles

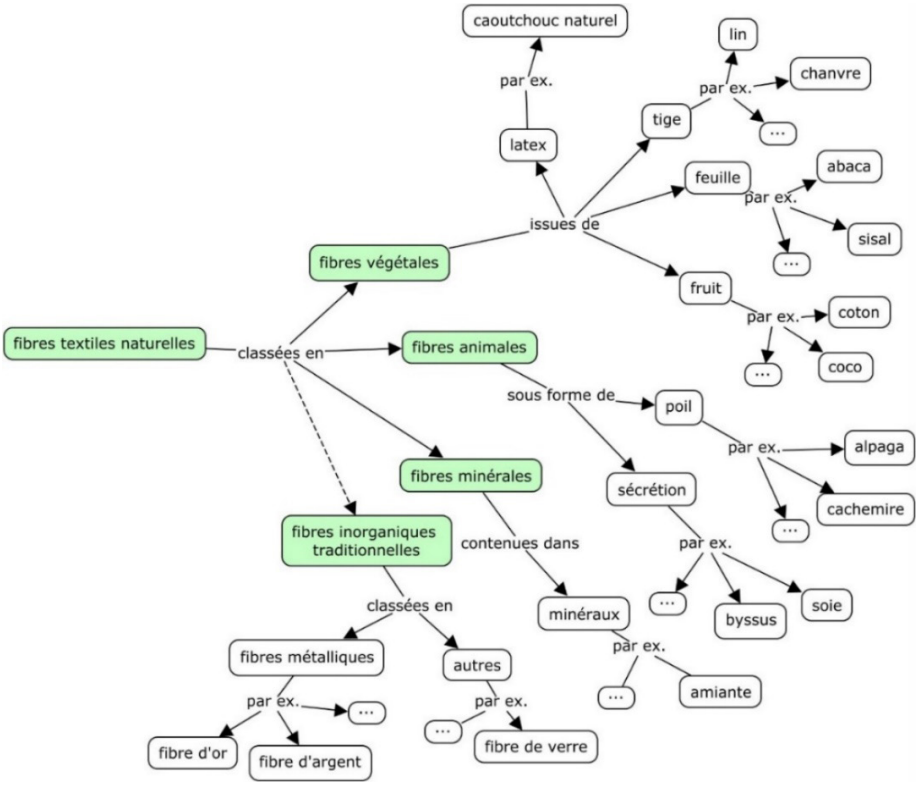


Fig. 2 – *Les fibres textiles jusqu'en 1884.*

Jusqu'à la fin du XIX^e^ siècle, les seuls matériaux utilisés dans la production textile étaient les *fibres naturelles*, qui se trouvent à l'état libre dans la nature. C'est le cas de trois catégories de fibres : *fibres végétales*, *fibres animales* et *fibres minérales*. Les fibres végétales sont distinguées selon la partie du corps végétal, qui les fournit en fibres provenant de la tige (ex. le lin), de la feuille (ex. l'abaca) et du fruit (ex. le coton). Un type particulier de fibre végétale est représenté par le caoutchouc naturel, obtenu à partir d'une substance liquide produite par les plantes comme l'hévéa. Les fibres animales sont constituées par les poils de certaines espèces d'animaux, tels que la chèvre cachemire, et par des substances sécrétées par des animaux particuliers, comme le ver à soie. Les fibres minérales font partie de la structure de certains minéraux, dont par exemple l'amiante. À part ces trois catégories classiques de fibres naturelles, il existe aussi une catégorie de transition que nous appelons *fibres inorganiques traditionnelles*, regroupant les fibres d'origine minérale qui sont obtenues depuis de longs siècles par un processus de transformation chimique, par ex. les fibres de verre (Baum et Boyeldieu 2018, 258) et les fibres d'or (Agulhon 1962, 123).

2.2.2. Depuis 1884 jusqu'à la fin des années 1930 : l'introduction des fibres artificielles

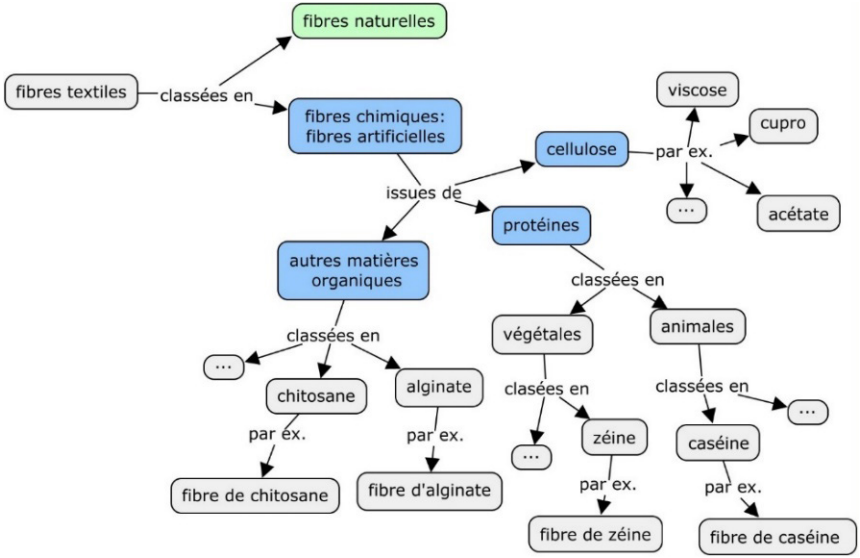


Fig. 3 – *Les fibres textiles depuis 1884 jusqu'à la fin des années 1930.*

La fin du XIX^e^ siècle est marquée par l'introduction des fibres chimiques appelées « soies artificielles », dont la production a été brevetée pour la première fois en 1884 par le comte Hilaire de Chardonnet (Agulhon 1962, 7-8). À la différence des fibres inorganiques traditionnelles, les soies artificielles constituent un premier substitut des fibres naturelles et c'est probablement pourquoi elles sont souvent considérées comme les premières fibres textiles manufacturées. L'introduction des soies artificielles a donné naissance à la catégorie de fibres appelée *fibres artificielles*, contenant des fibres obtenues par le traitement chimique des polymères naturels (Baum et Boyeldieu 2018, 257). Les substances naturelles traitées incluent la cellulose (par ex. la soie au cuivre ou cupro, brevetée en 1890 (Agulhon 1962, 9), et plus tard aussi les protéines végétales (par ex. la fibre de zéine, développée à la fin des années 1930 (Agulhon 1962, 14) et animales (par ex. la fibre de caséine, mise au point au début du XX^e^ siècle (Génin 1954, 521-522) et d'autres matières organiques, telles que les polysaccharides (par ex. la fibre d'alginate inventée dans les années 1910 (Miraftab *et al*. 2001, 165).

2.2.3. Depuis la fin des années 1930 jusqu'à la fin des années 1960 : l'introduction des fibres synthétiques et des nouvelles fibres inorganiques

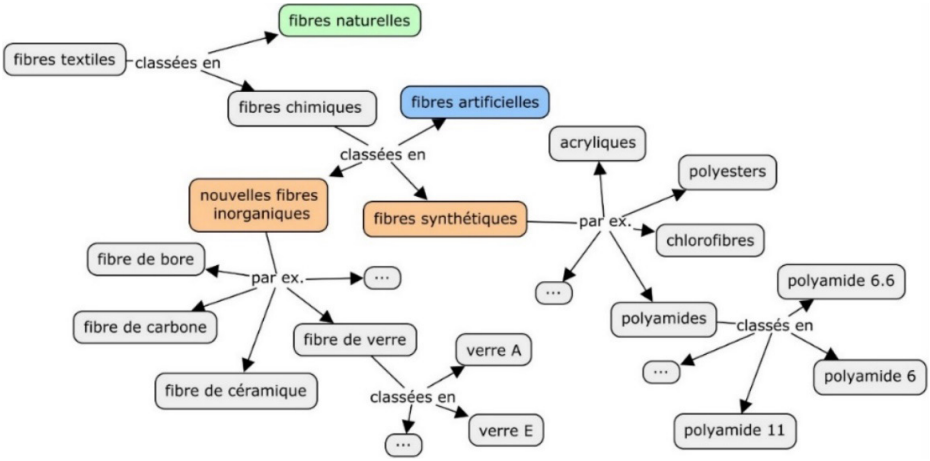


FIG. 4 – *Les fibres textiles depuis la fin des années 1930 jusqu'à la fin des années 1960.*

À la fin des années 1930, la production de textiles a connu un grand changement avec l'invention des fibres synthétiques, constituées par des polymères nouveaux fabriqués par synthèse chimique (Baum et Boyeldieu 2018, 257). Les premières fibres synthétiques, les polyamides, ont été lancés sur le marché en 1938, aux États-Unis sous le nom de *Nylon* (polyamide 6.6) et en Allemagne sous le nom de *Perlon L* (polyamide 6) (Agulhon 1962, 14-15). La fabrication de fibres synthétiques s'est développée considérablement après la Seconde Guerre mondiale. Actuellement, plusieurs familles de fibres synthétiques – par exemple, les polyesters, les acryliques, les polyuréthanes, les aramides et les chlorofibres – peuvent être distinguées[3] et, à l'intérieur de chaque famille, de nouvelles déclinaisons des fibres continuent à être proposées.

Dans la même période, la technologie de production de certaines fibres inorganiques (par ex. la fibre de verre) a été perfectionnée et de nouvelles matières inorganiques commencent à être utilisées pour la fabrication de fibres, telles que les fibres de carbone, développées dans les années 1950 (Weidmann 2010, 80) et les fibres de bore, produites depuis les années 1960 (Bunsell et Renard 2005, 39).

3 Voir BISFA (2017).

2.2.4. Depuis les années 1970 jusqu'à présent : l'introduction des fibres spéciales

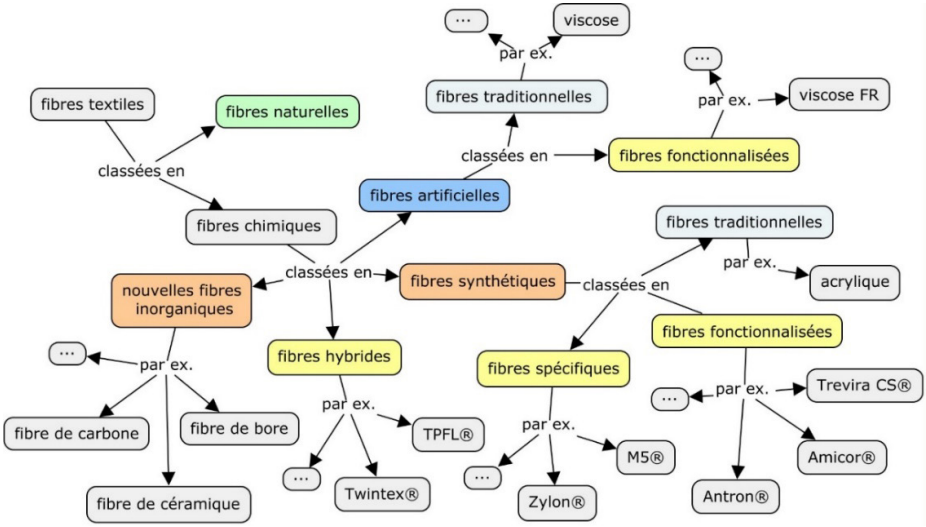


FIG. 5 – *Les fibres textiles depuis les années 1970 jusqu'à présent.*

Depuis les années 1970, trois nouvelles catégories de fibres se sont progressivement développées dans le cadre des fibres chimiques : les *fibres fonctionnalisées,* les *fibres spécifiques* et les *fibres hybrides*. Les fibres fonctionnalisées sont obtenues par un traitement chimique des fibres artificielles et synthétiques traditionnelles qui leur confère des propriétés fonctionnelles innovantes (Weidmann 2010, 16), telles que la résistance au feu (par ex. Trevira CS®, une fibre de polyester ignifuge (Weidmann 2010, 107) et la résistance aux micro-organismes (par ex. Amicor®, une fibre acrylique antimicrobienne (Sedelnik 2006, 120). Les fibres spécifiques représentent de nouvelles fibres synthétiques, qui se distinguent par des propriétés thermiques et mécaniques exceptionnelles. C'est le cas de la fibre PBO fabriquée par la société japonaise Toyobo et commercialisé sous la marque Zylon® (Weidmann 2010, 122). Les fibres hybrides constituent une catégorie particulière de fibres chimiques, parce qu'elles sont produites à partir des matières différentes, dont l'une sert de matrice et l'autre de renfort : le fil Twintex® fabriqué par l'entreprise Saint-

Gobain Vetrotex, est obtenu en utilisant le verre E et les fils thermoplastiques, comme le polypropylène (Weidmann 2010, 131).

2.3. Évolution des termes : *fibre*

L'analyse des termes figurant dans les ouvrages majeurs décrivant les fibres textiles depuis le XVIII^e^ siècle a révélé qu'avant la moitié du XIX^e^ siècle, il n'existait aucun terme particulier pour désigner les matières textiles en général. La langue française ne disposait que d'une série de termes désignant certains types de matières textiles :

- les fibres extraites de la tige et des feuilles : *fibre*, *filament*, *matière filamenteuse* ;
- les cotons et d'autres matières d'aspect semblable : *bourre végétale ;*
- les matières textiles secrétées par des espèces d'animaux particulières : *soie ;*
- les poils des bêtes à laines, telles que les moutons : *laine* ;
- les poils des autres animaux, tels que les chèvres : *poil*.

Les premiers termes désignant le concept de matière textile sont attestés depuis la deuxième moitié du XIX^e^ siècle : *matière textile* (Toustain 1859), *fibre* (*textile*) (Bezon 1863) et *textile* (attesté selon le Petit Robert 2018 depuis 1872).

À l'origine, le terme *fibre* a été utilisé dans le domaine de l'anatomie dans le sens de « formation d'aspect filamenteux, végétale ou animale ».[4] Entré dans le domaine du textile, le sens anatomique du terme s'est restreint pour désigner uniquement les longs filaments d'origine végétale, tirés de la tige ou des feuilles et utilisables dans la production de textiles. C'est le cas du terme *fibre* employé par Roland de La Platière (1785) pour désigner les filaments textiles d'abaca :

> « Dans le pays [aux Philippines], on appelle de ce nom les fibres préparées du figuier bannanier dont on vient de parler : on emploie ces fibres dans beaucoup de sortes de toileries & en cordages. On y voit des étoffes mélangées d'*abaca*, de soie & de coton : on en brode ; on en fait de la dentelle, &c. » (Roland de La Platière 1785, V)

4 Voir la première édition du *Dictionnaire de l'Académie française* (1694).

À partir de la moitié du XIX^e siècle, *fibre* – suivi ou non de l'adjectif *textile* – commence à être employé pour désigner une matière textile en général. Observons un extrait du *Dictionnaire général des tissus* de J. Bezon (1863) :

> « On désigne sous le nom de byssus une étoffe de filaments qui proviennent de certains mollusques. Celui de la *pinne-marine* est très long, très fin ; son moëlleux et son brillant lui donnent une grande ressemblance avec la soie. La *pinne-marine* est nommée *coquille porte-soie* par Aristote, qui signalait dans le byssus de ce mollusque une fibre textile. » (Bezon 1863, 401)

En ce qui concerne le terme *fibre* dans les dictionnaires, il faut mettre en évidence que *fibre* en tant que matière textile n'est enregistré dans les dictionnaires de langue qu'à partir de la deuxième moitié du XX^e siècle.[5] Dans le domaine textile, *fibre* désigne aussi un élément textile de courte longueur, par opposition aux termes *filament* ou *fil continu*, désignant des éléments textiles de grande longueur (Weidmann 2010, 283-284). Il s'agit d'un sens très technique qui ne figure pas, à l'heure actuelle, dans les dictionnaires de langue.

2.4. Évolution des termes : *artificiel*

Dans les dictionnaires de langue, *artificiel* est d'abord défini par opposition à *naturel* en tant que « qui se fait par art », dans le sens de « produit par l'homme » (par ex. Littré, 1873). Plus tard, les dictionnaires enregistrent d'autres acceptions, dont en premier lieu celle de « factice » ou bien « contraire à la nature » (à partir de la huitième édition du Dictionnaire de l'Académie, 1932) et ensuite d'autres acceptions dérivées, telles que « produit par la technique » et « créé par la pensée humaine » (Petit Robert, 2018).

Dans le domaine textile, le sens du terme *artificiel* évolue en fonction des changements significatifs dans l'organisation conceptuelle du domaine. Jusqu'à la mise au point des fibres chimiques, l'adjectif *artificiel* signifie « fait par l'homme à partir des fibres naturelles ». Toustain (1859) emploie le terme *tissu artificiel* dans le sens de « tissu produit par l'homme à partir des matières d'origine naturelle » et il le définit en opposition à *tissu naturel*, désignant les tissus animaux et végétaux, constitués de cellules :

5 Voir par ex. le *Dictionnaire alphabétique et analogique de la langue française : les mots et les associations d'idées* (1966).

> « Les tissus artificiels sont ceux que l'on exécute journellement de mille façons diverses, qui produisent ces variétés infinies qu'offre la fabrication, à l'aide de moyens variés, propres à transformer les matières premières en étoffes […]. » (Toustain 1859, 8)

Après l'invention des fibres chimiques à la fin du XIXe siècle, le sens de l'adjectif *artificiel* correspond à « fait par l'homme à partir des fibres manufacturées ». Tel est l'emploi dans cet extrait d'un manuel textile de H. Algoud (1912) :

> « Mais ces avantages sont fortement balancés par les défauts ou les insuffisantes qualités de cette nouvelle matière, dont on aurait pu penser un instant qu'elle allait concurrencer très sérieusement la soie naturelle. Lorsqu'il fut trouvé, ce fil artificiel était très inflammable […]. » (Algoud 1912, 18)

À cette époque, les *fibres artificielles* sont définies par rapport aux *fibres naturelles* en tant que fibres fabriquées à partir d'une matière qui ne se trouve pas à l'état libre dans la nature et elles couvrent l'ensemble des fibres manufacturées. Le nom de l'organisation BISFA qui s'occupe depuis 1928 de la standardisation des fibres manufacturées conserve une trace de cet état de l'évolution du domaine : « Bureau International pour la Standardisation des Fibres Artificielles », traduit en anglais par « International Bureau for the Standardisation of Man-Made Fibres ».[6]

Un autre changement du sens de *artificiel* est entraîné par le développement des fibres synthétiques à la fin des années 1930. Les fibres artificielles doivent être définies non seulement par rapport aux fibres naturelles, mais aussi par rapport aux fibres synthétiques, constituant une catégorie de fibres chimiques. Dorénavant, le terme *fibres artificielles* ne désigne plus toutes les fibres chimiques, mais seulement celles fabriquées à partir des polymères naturels.

3. Analyse du corpus

Les termes désignant les fibres chimiques en français ont été extraits manuellement à partir du corpus contenant quatre types de sources :

- des catalogues des salons professionnels : Première Vision Yarns (12-14 février 2019) [1], Première Vision Fabrics (12-14 février 2019) [2] ;

6 Voir BISFA. http://www.bisfa.org.

- un document institutionnel : DGE/UBIFRANCE, *Textiles Techniques. Le futur se tisse en France*, 2006, 24 p. [3] ;
- un ouvrage de vulgarisation : C. Fauque - S. Bramel, *Une seconde peau : fibres et textiles d'aujourd'hui*, Éditions Alternatives, Paris, 1999, 155 p. [4] ;
- un manuel technique : D. Weidmann, *Aide-mémoire textiles techniques*, Dunod, Paris, 2010, 294 p. [5].

Le corpus obtenu inclut 245 termes, constitués par les noms génériques et les noms de marque dans les proportions suivantes :

- 185 noms de marque (75 % du corpus) ;
- 60 noms génériques (25 % du corpus).

Il faut souligner que l'étude des dénominations des fibres chimiques est compliquée du fait que la plupart des termes ne sont pas enregistrés dans les ouvrages lexicographiques. Au total, seulement 30 % des noms génériques (19 termes) et un nombre très réduit de noms de marque (13 termes sur 185) sont enregistrés dans au moins un des dictionnaires de langue[7] : c'est le cas des termes génériques *viscose*, *acétate* et *cupro* et des noms de marques *Albène*, *Orlon* et *Kevlar*.

Notre analyse se propose les objectifs suivants :

- la définition des termes rassemblés ;
- leur répartition dans le réseau conceptuel du domaine ;
- le classement des synonymes et des équivalents en italien et en anglais.

Du point de vue linguistique, l'analyse porte principalement sur l'examen des procédés de formation des noms génériques en français. Les informations sur les termes sont résumées dans des fiches terminologiques, contenant les champs suivants : le terme en français et sa marque morphologique, la source du terme, le type de formation, le sous-domaine, la définition en français et sa source, les notes diachroniques, le synonyme et les équivalents en italien et en anglais. Des informations complémentaires sont fournies dans les notes explicatives et encyclopédiques, figurant sous la fiche terminologique. À titre illustratif, nous présentons la fiche contenant les informations sur le terme générique *cupro*.

7 Nous avons consulté cinq dictionnaires publiés après 1884 : *Dictionnaire de l'Académie française* (1935) ; *Dictionnaire alphabétique et analogique de la langue française : les mots et les associations d'idées* (1966) ; *Trésor de la langue française* (1971-1994) ; *Dictionnaire de l'Académie française* (1992, 2000, 2011) ; *Petit Robert* (2018).

Terme	*cupro*, n. m.
Source	[1]
Type de formation	Abréviation
Sous-domaine	Fibre artificielle
Définition	Fibre fabriquée à partir de la dissolution de la cellulose dans une solution cupro-ammoniacale.
Source	H. Agulhon, *Les textiles chimiques*, Presses Universitaires de France/ Que Sais-Je ? Paris, 1962, p. 9.
Notes diachroniques	- datation : 1933 - abrév. de *cuproammoniacal* - présence dans les dictionnaires de référence : PR (2018)
Synonyme	*rayonne cupro-ammoniacale*, n. f.
Code	CUP (ISO 2076 : 2013) CU (COMITEXTIL)
Équivalents	
IT	*cupro*, n. m. *raion cuprammoniacale*, n. m.
EN	*cupro* *cuprammonium rayon*

Fig. 6 – *La fiche terminologie de* cupro

4. Conclusion

Dans la période entre la fin du XIXe siècle et le début du XXIe siècle, la terminologie des fibres chimiques s'évolue en reflétant le degré de développement du domaine. La première phase, qui s'étend du milieu des années 1880 jusqu'à la fin des années 1930, est marquée par la production des fibres artificielles à partir des polymères naturels. En raison d'une diversification faible des produits et d'un taux de production relativement bas, le nombre de termes désignant les fibres chimiques à cette époque est assez réduit. À la fin des années 1930, l'invention des fibres synthétiques, dont la production n'est pas dépendante de l'existence d'un polymère naturel, entraîne une augmentation nette du nombre de producteurs, cherchant – surtout à partir des années 1950 – à mettre au point des fibres innovantes. Une partie essentielle du développement de nouveaux produits est constituée par la création de leurs

dénominations, car le succès d'un produit sur le marché est conditionné aussi par l'efficacité de son nom. Dans la deuxième moitié du XX^e siècle, on assiste ainsi à une prolifération des noms de marque, permettant aux fabricants de distinguer leurs produits de ceux de la concurrence. Il faut aussi remarquer qu'au début les noms de marque sont plutôt liés au marché national ou local, alors que la mondialisation de la production vers la fin du XX^e siècle a pour conséquence un affaiblissement du lien entre les noms de marque et le lieu de production.

Références

Agulhon, Henri. 1962. *Les textiles chimiques*. Paris : Presses Universitaires de France/ Que Sais-Je ?.

Algoud, Henri. 1912. *Grammaire des Arts de la soie*. Paris : Schemit.

Baum, Maggy et Chantal Boyeldieu. 2018. *Dictionnaire encyclopédique des textiles*. Paris : Eyrolles.

Bezon, Jean. 1863. *Dictionnaire général des tissus anciens et modernes*, t. VIII. Lyon : Imprimerie de Th. Lépagnez.

Bonetti, Ferruccio, Stefano Dotti et Giuseppe Tironi. 2012. *Fibre tessili. Struttura, caratteristiche, proprietà*, Milano : Tecniche Nuove.

Bunsell, Anthony R. et Jacques Renard. 2005. *Fundamentals of Fibre Reinforced Composite Materials*. Bristol et Philadelphia : IOP Publishing Ltd.

Bureau International pour la Standardisation des Fibres Artificielles. Consulté le 20 octobre 2019. http://www.bisfa.org.

CCSTI du Rhône, Université de Lyon. 2009. « Textiles d'hier, d'aujourd'hui et de demain. Dossier pédagogique. » Consulté le 20 octobre 2019. https://www.soierie-vivante.asso.fr/PDF/fildhier.pdf.

DGE/UBIFRANCE. 2006. « Textiles Techniques. Le futur se tisse en France. » Consulté le 20 octobre 2019. https://www.entreprises.gouv.fr/files/files/directions_services/secteurs-professionnels/etudes/textileF.pdf.

Fauque, Claude et Sophie Bramel. 1999. *Une seconde peau : fibres et textiles d'aujourd'hui*. Paris : Éditions Alternatives.

Génin, G. 1954. « Fibres de protéine. Caséine et arachide. » *Le Lait* 34 : 521-526.

International Bureau for the Standardisation of Man-Made Fibres. 2017. « Terminology of man-made fibres », Consulté le 20 octobre 2019. http://www.bisfa.org/wp-content/uploads/2018/06/2017-BISFA-Terminology-final.pdf.

Miraftab Mohsen, Q. Qiao, John F. Kennedy, Subhash C. Anand et Graham Collyer. 2001. « Advanced materials for wound dressings : biofunctional mixed carbohydrate polymers ». In *Medical Textiles. Proceedings of the 2nd international Conference, 24th and 25th August 1999, Bolton Institute, UK*, édité par Subhash C. Anand, 164-172. Abington : Woodhead Publishing Ltd.

Roland de La Platière, Jean-Marie. 1785. *Encyclopédie méthodique. Manufactures, arts et métiers*, t. I. Paris : Panckoucke.

Sedelnik, Natalia. 2006. « Health Properties of Polyacrylonitrile Blankets Made with the Anti-Microbial Fibre Amicor Plus ». *Fibres & textiles in Eastern Europe* 59 : 120-124.

Toustain (D'Elbeuf), Félix. 1859. *Nouveau manuel complet de la fabrication des tissus de toute espèce*. Paris : Librairie Encyclopédique de Roret.

Union des Industries Textiles. 2017. « Livre blanc sur les textiles intelligents ». Consulté le 20 octobre 2019. http://www.textile.fr/wp-content/uploads/2017/03/livre_blanc_UIT_2017_web.pdf.

Weidmann, Daniel. 2010. *Aide-mémoire Textiles techniques*. Paris : Dunod.

Abstract

In order to define specialty terms clearly and efficiently, it is important to have knowledge of the place occupied by the designated concepts in the structured network of relationships that constitute the conceptual organization of the domain. This article reconstructs the evolution of the field of chemical textile fibers from its inception in the late-nineteenth century to the end of the twentieth century, taking into account the terminology used. To begin, chemical fibers are defined and located in the textile industry. Thereafter, the conceptual evolution of the field of chemical fibers within that of textile fibers is traced and then illustrated through meaning changes in two key terms in the field : *fibre* and *artificiel*. Finally, the construction of a terminological resource containing terms designating chemical fibers is described, followed by an exemplification of the manner in which terminology sheets are constructed.

Eugen Wüster's Sign Typology – Some Observations

Marija Ivanović

Gymnasiumstraße 50
1190 Vienna
marija.ivanovic@univie.ac.at

Abstract. Eugen Wüster was the founding father of what became known as the Vienna School of Terminology. With the posthumously published General Theory of Terminology (Wüster 1979), he developed a meta-theory for terminology. In its chapter on signs, he broadened the concept of "term" (as later defined in ISO 704 2009): it now encompasses all forms of signs. The General Theory of Terminology also encompasses a draft sign typology, which is the basis of yet unfinished DIN 2338. This paper looks at Wüster's sign typology, which is structured as a divisor-combinatory concept system, and tries to find out whether it is complete and coherent. As a guideline for the analysis it uses the rules which were the basis for the construction of this concept system. Finally, it provides a perspective for a possible application of sign typologies.

1. Introduction

Eugen Wüster was the founding father of what is known as the Vienna School of Terminology. First this article reviews at Eugen Wüster's draft for a sign typology, which encompasses signs appealing to all senses.

Over the last decades, different forms of non-verbal representation have gained importance in terminology science and terminology work. This can be seen in the work of Galinski and Picht (1997), Picht (1999), Madsen (2013), Lervad *et al.* (2013), as well as in terminology standards such as ISO 704 (2009) or ÖNORM A 2704 (2015). These approaches usually focus on static visual signs. ÖNORM A 2704 (2015, 47) mentions the relevance of different media for terminology work including animations and acoustic material. While most approaches deal with different possibilities of visual representa-

tion of concepts, it is also suggested that other forms of representations could be useful for future applications such as virtual reality, as stated by Galinski and Picht (1997, 58).

Eugen Wüster's sign typology was an early attempt at providing a semiotic approach encompassing all senses. Its starting point was Wüster's meta-theory of terminology, which was published posthumously as the General Theory of Terminology in 1979 (Wüster 1979)

In this paper, I will analyse whether the first part of Wüster's sign typology is complete and coherent on the highest level and tries to find out whether and where it shows inconsistencies. As a methodological tool, a self-reflective terminological approach is used. The rules for the construction of the concept system, which is the basis for Wüster's sign typology, serve as guidelines.

In the first part of this paper, Eugen Wüster and his sign typology are introduced. Then the theoretical considerations on concepts and the specific concept system on which Wüster's sign typology is based are provided. In the next part, these theoretical considerations are used to examine the strengths and weaknesses of Eugen Wüster's sign typology. The last part includes a perspective for the use of sign typologies in future applications.

As the basis for the translation and terminology of the General Theory of Terminology (GTT) and sign typology for this paper, I will use its unpublished Translation by Charles Gilreath.

2. Eugen Wüster

Eugen Wüster was an Austrian electrical engineer and owner of a machine tool factory. As a pupil, he had already compiled and edited a four-volume encyclopaedic Esperanto dictionary. Even at this young age, he had a strong interest in planned languages, which evolved into an interest in language planning. This interest is also visible in Wüster's doctoral thesis on International Standardization of Technical Languages, particularly in Electrical Engineering (1931). It became an important work in applied linguistics. For Wüster himself this thesis was the beginning of a life-long commitment to standardization and terminology in a variety of fields. (Lang 1998 :13)

Throughout his life, Wüster worked as a self-taught linguist and was an active member of the Technical Committee ISA 37 (International Standards Association) which dealt with terminology standardization before World War II. After the war, Wüster committed himself to bringing the TC 37 ("Terminology : Principles and Coordination") at the now International

Standardization Organization (ISO) back to life and continued working for it (Lang 1998, 14-17).

He also worked in the fields of bibliography, lexicography, orthography, and contributed to the development of the Universal Decimal Classification (Lang 1998).

From 1972 onwards, Wüster was professor of lexicology, lexicography and terminology science at the Institute of Linguistics of the University of Vienna. His life-long interest in standardization and terminology was the basis for the formation of meta-rules for terminology work. The manuscript he prepared for his lectures was edited and published by Helmut Felber as the General Theory of Terminology (Lang 1998, 21) .

3. Wüster's semiotic approach

One of the basic models of terminology science is a version of the semiotic triangle with its three constituting parts object, concept and designation. This model was also the starting point for Wüster's sign typology.

In his GTT Wüster (1991, 59f.) decided to expand the concept of the hitherto mainly verbal form of designation so it could encompass signs of all types. For Wüster this was necessary because a variety of different kinds of signs is used to represent concepts in technical languages. The expansion of the concept designation also made an expansion of linguistics towards a semiotic perspective necessary, as he had seen in the work of Saussure.

A designation for Wüster could now encompass signs which appeal to all the senses. Wüster tried to structure the different sign types in a divisory table of characteristics carriers. Divisory tables of characteristics carriers are one way to structure a concept system, which is presented in the GTT.

Wüster developed this sign typology because he was sure that "If we do wish to say something about the various types of signs, then we can only do so if we first develop a classification of signs. What we have to say about signs is included in the classification" (Wüster, n.d., 5, ch 6)[1]

Wüster's sign typology consists of two parts. The first part of the table includes the categories of characteristics carriers which are used to describe signs in general :

1 The translation of the GTT by Gilreath does not have continuous page numbers, therefore the relevant chapters are also provided.

a) Connection between sign and object
b) Sense organs
c) Structure
d) Relationship to language
e) Nature of referent
f) Frequency, importance
g) Specialization
h) Transparency
i) Directness of meaning

Each category encompasses several characteristics carriers. The characteristics carriers themselves are also concepts.

The second part of the table are the characteristics categories j) – l) which are used to describe specific types of signs, namely written signs, a sub-form of visual signs. The observations in this contribution focus on categories a) – i) of the sign typology and are a first attempt at coming closer to what Wüster was aiming at.

The sign typology was a draft for DIN 2338 (1971), which was never finished. Therefore, the analysis of his draft for a sign typology must be conducted carefully in order to avoid jumping to conclusions. For the time being, it can only consist of observations and reflections on what can be found in the unfinished state of the sign typology. To gain a profound understanding of Wüster's sign typology, future work will make an analysis of the influences necessary, as well as an analysis of the unfinished DIN 2338.

4. Concepts, characteristics and divisory tables of characteristics carriers

Terminological work is usually concerned with the analysis of concepts and concept systems. In Wüster's words a concept

> " apart from "concepts" of individual objects, is the common element perceived in several objects and used as a means of classifying one's thoughts ("conceiving") and, therefore, as a means of communication. Hence, a concept is an element of thought." (Wüster, n.d., 8f., ch 3).

4.1. Characteristics

In terminology theory, concepts are identified by their characteristics. Characteristics are abstracted properties of the objects that are to be mentally represented. In terminology work these characteristics are then used to distinguish, define, describe and classify concepts as well as to structure concepts

systems (Wüster 1991, 16; Weissenhofer 1995, 7; Arntz and Picht 1995, 54ff.; Felber and Budin 1989, 70). The structure of the subject field (Wüster 1991, 16) as well as the purpose of the terminology work determine which characteristics should be considered. Different approaches to a certain subject field make a focus on different properties of an object and thereby of characteristics of a concept necessary (Felber and Budin 1989, 69).

For this paper, Wüster's basic approach to analysing concepts and concept systems is used to analyse his sign typology.

Wüster (Wüster 1991, 8f.) distinguished between the intension and the extension of a concept: while the extension of a concept encompasses all the subordinate concepts, "[t]he intension of the concept is the aggregate of all characteristics which constitute a concept" (Wüster, n.d., 8f., ch 3). Wüster used the characteristics to delimit concepts from each other and structure concept systems (Wüster 1991, 9-15).

As a guideline for the observations on Wüster's sign typology only the characteristics of the concepts, which were relevant for the construction of the concept system are considered, as they are the basis for the relations within the concepts system.

4.2. Divisory-combinatory tables of characteristics carriers and their concept relations

Wüster used a specific concept system to structure his sign typology. The sign typology is a divisory-combinatory table of characteristics carriers. The ontological necessity for this sort of concept system and diagram is the possibility to combine all categories of characteristics carriers with each other (Wüster, n.d., 23ff., ch 3). In the case of Wüster's sign typology this means that from each of the categories a) – i) one characteristics carrier can always be combined with one of the other characteristics carriers from the other categories.

Within the categories, the concepts on the first level are logically coordinate to each other and show certain common characteristics:

"Frequently both of the concepts compared possess at least one additional feature, besides their common intension, which distinguishes them from each other. If the distinguishing features are of the same kind (i.e. if they belong to the same generic concept) then one speaks of logical coordination." (Wüster, n.d., 10, ch 3)

The characteristics carriers, which are concepts themselves, should therefore have certain characteristics in common and show characteristics which differentiate them from other characteristics carriers within the concept system.

As a divisory-combinatory concept system it has a divisory aspect as well. It is also based on the relation of logical subordination: within each column the characteristics carriers can be further subdivided to show subordinate concepts (Wüster, n.d., 25, ch 3).

Wüster describes the relation the characteristics of the concepts should have when a concept is logically subordinate to another as follows:

> "If one concept possesses all the characteristics of another plus at least one additional characteristic, then it is considered to be a specific (or subordinate) concept of the other, and the broader one is called the generic (or superordinate) concept. This kind of relation is called logical subordination (generic-specific relation) and logical superordination (specific- generic relation)." (Wüster, n.d., 10, ch 3)

These two forms of concept relations determine the structure of the concept system and the position of each concept. Looking at the structure of characteristics within the categories a) – i) reveals a hint to find unfinished aspects and possible inconsistencies within the sign typology as well as a door to questions for future developments.

5. Some observations on Wüster's sign typology

Wüster structures the sign typology according to four aspects of the sign: 5.1. Cause and Effect, 5.2. Sign Form, 5.3. Meaning, 5.4. Subject field and 5.6. Sign-meaning assignment. These aspects consist of the categories a) – i) which represent the characteristics categories. These categories will be analysed in this paper.

It is not the aim of this paper to argue whether the categories of the sign typology Wüster chose are complete or if they could be argued about. As they probably grew under the influence of different linguistic and semiotic theories the tracing of these influences is a task left to future work.

5.1. Cause and effect

This aspect encompasses category a), which has no title. Here Wüster distinguishes signs based on the connection between sign and object.

He distinguishes between 1. Natural signs and 2. Conventional signs. Natural signs hint at their object, such as when smoke is representing fire. (Wüster, n.d., 3, ch 6/7). In the case of 2. Conventional signs, the connection between the sign, say the word chair and its object, is based on a social agreement. The relationship between the object and its sign is arbitrary : the word chair as a sign does not show any of the properties the object chair has.

As Myking (2001, 46) and Järvi (1997, 70) point out that this is a distinction which can also be found in Peirce. Peirce (1998, 27) also knows "*likenesses,* or icons ; which serve to convey ideas of the things they represent simply by imitating them." These signs resemble their object in one way or another. For example, a footprint in the sand as a sign for a human who passed this way earlier.

If Peirce influenced Wüster, it would be interesting to find out why Wüster did not include this third sign type in category a).

In the following categories b) – i) Wüster restricts his analysis to conventional signs, because he considers them the basis for linguistic development (Wüster 1991, 62).

5.2. Sign form

When it comes to the aspect of sign form Wüster (1991, 62-68) differentiates two physical aspects of the sign : on the one hand category b), in which he analyses to which sense organs a sign appeals and category c), which refers to the structure in which a sign manifests itself.

5.2.1. Category b) Sense organs

When looking at category b) Sense organs, one would expect that Wüster organised the characteristic carriers of the category b) according to the involved sense organs. However, the sense organ to which they appeal is not the only relevant characteristic for the structure. The structure is also based on the suitability of the sign to stimulate the sense organs and the possibility to reproduce it (Wüster 1991, 63).

Only 1. Visual signs and 2. Auditory signs are structured according to the sense organ they refer to. For example, visual signs appeal to the eye in case of a traffic sign, or to the ear in case of a whistle.

3. Sensory contact signs encompass olfactory signs which affect the nose – like the warning smell of gas can. But this category also encompasses tactile signs which – in the form of braille – are essential for blind persons and appeal to the sense of touch. For Wüster this last group is of utmost importance for these persons, but not often used by others.

Category 4. Object signs does not appeal to one single sense either. Signs in this category, according to Wüster, are signs as a whole, which are three-dimensional objects and would probably appeal to several senses. Wüster mentions a loaf of bread outside a bakery or a nod of the head (Wüster 1991, 66). The loaf of bread is a visual sign but can also be touched, and maybe even has a distinct smell. The nod of a head is certainly a visual sign but by touching the head can also be felt.

The characteristics carriers in category b) Sense organs do not refer to one sense organ each and therefore do not completely adhere to the rules set by the structure Wüster chose for his concept system. The characteristics carriers as coordinate concepts do not have each one necessary specifying characteristic of the same kind. While traffic lights are perceived by the eye and auditory signs by the ear, the classification of olfactory signs/smells and tactile signs (which can be felt) within one characteristic carrier seems not completely finished.

Furthermore, 4. Object signs like a loaf of bread or gestures like the nod of a head, are relevant for a semiotic analysis, but do not address only one sense organ.

5.2.2. Category c) Structure

In this category, Wüster tries to analyse the possible structures of signs. On the level of his sign typology, he distinguishes between 1. Elementary sings and 2. Complex signs (= signs combinations). He does not define 1. Elementary signs in the text of the GTT but says that all signs which are not complex signs are elementary signs (Wüster 1991, 66). Therefore, the only assumption that can be made is that the signs body consists only of a single sign and is not constructed from several sign parts. The use of a single letter could be an example for an elementary sign. A complex sign on the other hand could be any word as it is built from several letters (Wüster 1991, 67), or

a complex graphic sign : such as the instructions one often finds when buying new furniture.

The specifying characteristic within this category is the structure of the sign. In complex signs the parts of the sign furthermore contribute to the meaning of the sign (Wüster 1991, 66f.).

Still, it is not clear what the basic elements are which build an elementary sign. It might be assumed that whether or not the sign parts contribute to the meaning might also be relevant for this aspect. This could hint to another not completely elaborated aspect of the typology.

The difference between the text of the GTT and the draft of the sign typology could be due to the fact that the GTT was published posthumously and is based on the manuscript of Wüster's lecture.

5.3. Meaning

When it comes to the analysis of the way a sign carries meaning, Wüster distinguishes two categories of characteristics carriers : category d) Relationship with language and category e) Nature of referent, which looks at the possible objects of a sign.

5.3.1. Category d) – Relationship to language

The analysis of category d) is difficult without knowing why Wüster chose these sign types as characteristic carriers.

1. Phonetic signs according to Wüster represent certain language sounds and have an indirect connection to the meaning of the sign. Wüster mentions flashing signals used as Morse code. The verbal meaning of these signs manifests itself only through the Morse code. While 2. Semantic signs directly refer to the meaning of a sign : "these are conventional signs which, instead of being assigned to a phonetic meaning, are assigned to a non-phonetic concept (or individual)" (Wüster, n.d., 6/14f.). Semantic signs are represented in a way which can easily be understood without knowing the language. Examples are graphic signs or names (Wüster 1991, 71)

Distinctions like these seem to be influenced by other authors so that I cannot extrapolate the concept relations without analysing the influences and aims they were based on.

The third category on this level are 3. *Gedankenzeichen* a working translation could be *signs of thought*. These signs are not defined in the GTT and therefore not translated by Gilreath.

It would also be interesting why Wüster makes the distinction between 1. Phonetic signs, 2. Semantic signs and 3. Signs of thought (*Gedankenzeichen)*. The distinction does not seem to be completely elaborated. Wüster analyses the relationships between sign and language. It would therefore also be necessary to find out why he chose these three characteristics carriers and how the unfinished aspects were developed further in the unfinished DIN 2338 (1971).

5.3.2. Category e) Nature of Referent

Wüster also analyses signs according to the nature of the object they refer to. To Wüster, this category is only relevant for conceptual signs (Wüster 1991, 72). Wüster probably decided to limit this category to conceptual sign because he wanted to focus on aspects of the signs relevant to a terminological analysis. Still, signs, which are not conceptual signs, do have a referent, which could also be considered in a sign typology.

A look at the distinction Wüster does draw is still interesting and might be expanded to all signs.

Wüster distinguishes in category e) between 1. Signs for sensory impressions, 2. Signs for things and 3. Signs for non-material referents, but it is not always clear where the line between signs for sensory impressions, signs for things and signs for non-material referents is drawn.

When speaking about 2. Thing signs and 3. Non-material referents, Wüster (1991, 73) mentions the sign for the Red Cross, as referring to a non-material object, and a nation state, as a material object. The question what is material and what is not seems not completely solved here. A nation state as well as the Red Cross are on the one hand referring to certain values or beliefs, such as compassion or the feeling of belonging to a nation, which are quite abstract. At the same time the Red Cross as well as each nation state manifest themselves in tangible structures, be that buildings or borders.

Thoughts might be another example of everyday non-material referents, which can turn into material referents. Thoughts can be measured or at least traced by functional magnetic resonance imaging when the neural activity is traced by analysing the cerebral blood flow it causes : and as they can be traced, have they not become something material ? The question as to what

is material and what is not could depend on the perspective from which one looks at it.

5.4. Classification by Subject field

According to Wüster "The subject field provides us with two types of characteristics : on the one hand the role of signs in the system of signs, and, on the other hand, the attribution of signs to their subject fields." (Wüster, n.d., 17 ch 6) and therefore there are once more two categories which are relevant for his signs typology.

5.4.1. Category f) – Frequency, Importance

In category f) Wüster (1991, 73f.) distinguishes signs by their role and importance within the signs system in 1. Basic signs which are frequently used, and form the basis of a sign system, like basic writing signs. The second category are 2. Special signs – which are used for special purposes.

The question I ponder upon is whether the special signs, Wüster (n.d., 25 ch 7) later mentions punctuation marks in writing or word-division signs like hyphens, really are special signs or whether they are an essential part of the sign system, as well. Especially when it comes to signs which are not writing signs, such as the signs used in architectural drawings.

5.4.2. Category g) Specialization

In category g) Specialization the line between Wüster's distinction in 1. Non-technical and 2. Technical signs is difficult to draw.

As an example for non-technical signs, Wüster mentions letters of the alphabet or a nodding of the head. Examples for technical signs for him are mathematical signs, traffic signs or electro-technical signs (Wüster 1991, 74).

While it is obvious that these signs are part of certain subject fields, some of these signs also pervade everyday life. It is therefore difficult to strictly classify them in non-technical and technical signs. What would a city be without traffic signs and how would we manage our everyday lives without knowing about mathematics ? And on the other hand, technical signs could also be composed of letters which are, according to Wüster, non-technical signs.

5.5. Sign-meaning assignment

Wüster analyses the connection between the form of the sign and its meaning. Two categories are relevant for Wüster. Category h) Transparency and category i) Directness of meaning, whether the sign is used in its first or second meaning.

5.5.1. Category h) – Transparency

In category h) Wüster analyses the level of similarity between object and sign. In this category Wüster first distinguishes between imitative signs and non-illustrative written sings.

Imitative signs show certain properties of the represented object, and in this way are a form of copy. But, according to Wüster, they are a copy which is extremely simplified and their interpretation is based on social convention (Wüster 1991, 75). Wüster does not limit this category to any specific kind of sign.

The distinction between 1. Imitative sings in general and 2. Non-illustrative written signs on the same level seems to distinguish between sings in general (which can be sounds as well as a variety of visual signs) and a subcategory of written signs, which for Wüster (1991, 78) are always visual signs.

The question which arises for me is whether signs which are not written signs, like a sound, can also represent their object without being any kind of copy of it? Such as the spoken word for bird does not show any characteristics of the object bird.

5.5.2. Category i) Directness of meaning

In category i) Wüster distinguishes between 1. Primary signs, which are used in their primary meaning, for example a letter for a sound like A as the sign for a sound, and 2. Secondary signs. The same sign A here would not be a sign for a sound, but could be the sign for Ampere in physics (Wüster 1991, 77).

This is a distinction which might be problematic. It could sometimes be difficult to define when a signs is used as a primary sign. Such as in the case of arrows: they can be vectors in mathematics, show a certain direction as traffic signs and can show that something is being concluded in a formalized way of writing.

6. Conclusion and Perspective

From my perspective, there are several unfinished and not completely structured aspects in Wüster's sign typology. Which are certainly due to the fact that the sign typology was never completed. For future work it would be interesting to analyse based on which influences these structures developed and how they were later perceived, for example in the version of the pre-norm of DIN 2338 (1971). Another possible approach would be to look into Wüster's notes on the GGT in the Wüster Archive at the University of Vienna.

However, apart from the analysis of the work of one of the pioneers of terminology science, the question remains: what could a sign typology be used for?

As the subject of these proceedings is Terminology and Ontologies – Theories and Applications I want to try to provide a connection between different signs types and their communication using the ontology-markup language XML by offering an example from research.

XML is one of the web- based ontology languages or ontology- markup languages as stated by Gómez- Pérez *et al*. (20043, 201) which are used "in the exchange of a wide variety of data on the Web and elsewhere." (W3C 2019)

This could include a variety of forms: for example, the characteristics of certain verbal and non-verbal signs. The work of Ranasinghe *et al*. (2012) hints in this direction: to "remotely and digitally communicate the sensation of taste" Ranasinghe *et al*. (2012, 409) developed on the basis of XML a format named TasteXML and used it to exchange information about taste sensations between a mobile application and a receiver over the Internet. The receiver is a digital taste simulator and reconstructs the taste sensations based on the sent information.

This is an example of the formulation of a taste in TasteXML for the taste *bitter*, as presented in Ranasinghe *et al*. (2012, 411):

<?xml version="1.0" encoding="UTF-8"?>

<TasteXml>

<type>request</type>

<method>SendTasteToFriend</method>

<information>

<ID>T0001</ID>

<FriendID>4</FriendID>

<FriendName>Daniel Jackson</FriendName>

<TasteType>basic</TasteType>

<taste intensity="70">Bitter</taste>

</information>

</TasteXml>

TasteXML has three basic taste sensations and is able to describe the intensity of tastes by using three levels. According to Ranasinghe *et al*. (2012, 411) the format can easily be expanded to describe more complex sensations, as well as several sensations occurring at once. It can furthermore be adapted to the needs of social networks.

In terminology science Galinski and Picht (1997, 58) envisioned the possibilities non-verbal sings offer for virtual reality.

Since then, electronic media have become an important tool in terminology work, as ÖNORM A 2704 mentions (2015, 47), and also in everyday communication. People mainly use text, sound, image and video or combinations to communicate. However, as the possibility of virtual reality suggests, real-life human interaction usually involves all senses as well as gestures and interaction with objects.

It is therefore easy to connect the vision of Galinski and Picht with the possibilities Ranasinghe *et al*. (2012, 410) see for taste sensations:

> "Being able to communicate taste sensations digitally has distinct advantages in several domains including multisensory communication, mixed and virtual reality, and entertainment. Virtual interaction can be enhanced through this technology by integrating the sense of taste in the virtual environment."

According to Ranasinghe *et al*. (2012, 415) there is also the possibility to use this approach in internet marketing, online shopping or the sharing of tastes in digital social networks.

These scenarios are not limited exclusively to taste sensations. Other forms of sensations and sign could also be communicated using some form of ontology mark-up language.

Categories of characteristics									Classification of written signs (b 1.3.1.1)			
Cause and effect	Sign form		Meaning		Subject field		Sign-meaning assignment		Combination of categories of characteristics			
	Sense organs	Structure	Relationship to language	Nature of referent	Frequency, importance	Specialization	Transparency	Directness of meaning	c, f	d, h, i		i, d
a	b	c	d	e	f	g	h	i	j	k		l
1. Natural signs 2. Conventional signs 2.1. Darstellende Zeichen 2.1.1. Nennzeichen	1. Visual signs 1.1. Light signs 1.2. Color signs 1.3. Gestaltzeichen 1.3.1. Form signs .... 1.3.2. Size signs 1.4. Combined visual signs 2. Auditory signs 3. Sensory contact signs 4. Object signs	1. Elementary signs 1.0.1. Zeichen mit einfachem Körper 1.0.2. Zeichen mit zusammengesetztem Körper .... 1.1. Autonomous elementary signs 1.2. Complemen tary signs 1.2.1. Supplemen tary signs 1.2.2. Structural signs 2. Complex signs	1. Phonetic signs 1.1. Segmental phonetic signs 1.2. Suprasegmental phonetic signs 1.2.1. Stress marks 1.2.2. Pause marks 2. Semantic signs 2.1. Representational semantic signs 2.1.1. Conceptual signs .... 2.1.2. Situational signs .... 2.2. Influence signs 2.2.1. Anweisungszeichen 3. Gedankenzeichen	1. Sensory impression signs 1.1. Visual impression signs 1.1.1. Signs for color 1.2. Acoustic signs 2. Thing signs 2.1. Schaltzeichen 3. Signs for nonmaterial referents 3.1. Zahlzeichen	1. Basic signs 2. Special signs	1. Nontechnical signs 2. Technical signs S e l e c t i o n : 2.1. Signs used in calculation 2.2. Traffic signs 2.3. Electrotechnical signs	1. Imitative signs 1.1. Illustrative signs 1.1.1. Direct representation 1.1.2. Indirect representation 2. Nonillustrative written signs	1. Primary signs 2. Secondary signs	1. Elementary written signs 1.1. Occasional written signs 1.2. Writing signs 1.2.1. Basic writing signs 1.2.2. Special writing signs .... 2. Complex written signs	1. Phonetic written signs 1.0.1. Phonetic signs 1.1. Figurative phonetic written signs 1.2. Nonfigurative phonetic signs 1.2.1. Primary nonfigurative phonetic written signs 2. Semantic written signs 2.0. Pictorial signs 2.1. Figurative semantic written signs 2.1.1. Primary figurative semantic written signs	2.1.2. Secondary figurative semantic written signs .... 2.2. Nonfigurative semantic written signs 2.2.1. Primary nonfigurative semantic written signs 2.2.2. Secondary nonfigurative semantic written signs	1. Primary written signs 1.1. Primary phonetic written signs 1.2. Primary semantic written signs 2. Secondary written signs 2.1. Secondary phonetic written signs 2.1.1. Rebus signs 2.2. Secondary written signs 2.2.1. Buchstabenzeichen
							Legend: ..... } = omitted subordinate concepts ..... (with 4 or 5 digits)					

TAB1 – Wüster's sign typology in English, based on the translation of the GTT by Charles Gilreath. Charateristics carriers which are not defined in the GGT and/or not translated are coded in blue.

References

Arntz, Reiner, and Heribert Picht. 1995. *Einführung in die Terminologiearbeit*. 3rd ed. Hildesheim, Zürich, New York : Georg Olms.

DIN 2338. 1971. 'Begriffssystem Zeichen'. Beuth.

Felber, Helmut, and Gerhard Budin. 1989. *Terminologie in Theorie und Praxis*. Tübingen : Narr.

Galinski, Christian, and Heribert Picht. 1997. 'Graphic and Other Semiotic Forms of Knowledge Representation in Terminology Management'. In *Handbook of Terminology Management*, edited by Sue Ellen Wright and Gerhard Budin, 1 :42-61. Amsterdam/Philadelphia : Benjamins.

Gómez-Pérez, Asunción, Mariano Fernández-López, and Oscar Corcho. 20043. *Ontological Engineering : With Examples from the Areas of Knowledge Management, e-Commerce and the Semantic Web.*

ISO 704. 2009. 'Terminology Work - Principles and Methods'.

Järvi, Outi. 1997. 'The Sign Theories of Eugen Wüster and Charles S. Peirce as Tools in Research of Graphical Computer User Interfaces'. *Terminology Science & Research* 8 (1): 63-72.

Lang, Friedrich. 1998. 'Eugen Wüster _ His Life and Work until 1963'. In *Eugen Wüster. His Life and Work. An Austrian Pioneer of the Information Society.*, 13-26. Wien : TermNet.

Lervad, Susanne, Peder Flemestad, and Lotte Weilgaard Christensen. 2013. 'Introduction to Verbal and Nonverbal Representation in Terminology'. In *Verbal and Nonverbal Representation in Terminology. Proceedings of the TOTh Workshop 2013. Copenhagen, 8 November 2013*, edited by Susanne Lervad, Peder Flemestad, and Lotte Weilgaard Christensen, xi–xvi. Kopenhagen : DNRF's Centre for Textile Research & Institut Porphyre, Savoir et Connaissance.

Madsen, Bodil Nistrup. 2013. 'The Use of Linguistic and Non-Linguistic Data in a Terminology and Knowledge Bank'. In *Verbal and Nonverbal Representation in Terminology. Proceedings of the TOTh Workshop 2013. Copenhagen, 8 November 2013*, edited by Susanne Lervad, Peder Flemestad, and Lotte Weilgaard Christensen, 1-22. Kopenhagen : DNRF's Centre for Textile Research & Institut Porphyre, Savoir et Connaissance.

Myking, Johan. 2001. 'Sign Models in Terminology : Tendencies and Functions'. *LSP & Professional Communication* 1 : 45-62.

ÖNORM A 2704. 2015. 'Terminologiearbeit - Grundsätze und Methoden'.

Peirce, Charles S., and Peirce Edition Project. 1998. *The Essential Peirce, Volume 2 : Selected Philosophical Writings (1893-1913).* The Essential

Peirce. Bloomington, Indiana : Indiana University Press. http://search.ebscohost.com/login.aspx ?direct=true&db=nlebk&AN=458440&site=e-host-live.

Picht, Heribert. 1999. 'Einige Überlegungen zu nicht-sprachlichen Repräsentationen von Gegenständen und Begriffen'. *SYNAPS - A Journal of Professional Communication* 3 : 1-50.

Ranasinghe, Nimesha, Adrian Cheok, and Ryohei Nakatsu. 2012. 'Taste/IP : The Sensation of Taste for Digital Communication'. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, 409-416. ICMI '12. ACM.

W3C. 2019. 'Extensible Markup Language (XML)'. Accessed August 29. https ://www.w3.org/XML/.

Weissenhofer, Peter. 1995. *Conceptology in Terminology Theory, Semantics and Word-Formation : A Morpho-Conceptually Based Approach to Classification as Exemplified by the English Baseball Terminology*. Wien : TermNet.

Wüster, Eugen. 1931. 'Internationale Sprachnormung in der Technik, besonders in der Elektrotechnik'. Berlin : VDI-Verl. https ://ubdata.univie.ac.at/AC04202072.

—. 1979. *Einführung in Die Allgemeine Terminologielehre und Terminologische Lexikographie : 1 : Textteil*. Schriftenreihe Der Technischen Universität Wien. Wien [u.a.]: Springer in Komm.

—. 1991. *Einführung in die Allgemeine Terminologielehre und Terminologische Lexikographie*. 3rd ed. Bonn : Romanistischer Verlag.

—. o. J. „General Theory of Terminology and Terminological Lexicography". Übersetzt von Charles Gilreath.

Résumé

Eugen Wüster est le père fondateur de ce qui fut nommé plus tard la terminologie de l'École de Vienne. Dans son ouvrage posthume *La théorie générale de la terminologie* (1974), il développa une méta-théorie de la terminologie. Dans son chapitre sur les signes, il élargit le concept de « terme » (défini par la suite dans l'ISO 704 2009) : désormais, il engloberait toutes les formes de signes. *La théorie générale de la terminologie* inclut également une ébauche de typologie des signes, qui constitua la base du DIN 2338, à ce jour inachevé. Ce travail examinera la typologie des signes de Wüster, structurée sous forme d'un système conceptuel diviseur-combinatoire, et cherchera à déterminer si elle est complète et cohérente. En guise de fil conducteur de cette analyse, il

s'appuiera sur les règles, qui constituèrent la base de la construction de ce type de système conceptuel. Enfin, il livrera une perspective d'application potentielle des typologies des signes.

Vers une ontologie de la nomination et de la référence dédiée à l'annotation des textes

Agata Jackiewicz*, Nadia Bebeshina*, Manon Cassier*** **** Francesca Frontini*, Anais Halftermeyer**, Julien Longhi***, Giancarlo Luxardo*, Damien Nouvel****

*PRAXILING, Route de Mende 34199 Montpellier cedex 5
prénom.nom@univ-montp3.fr
http://www.praxiling.fr
**Laboratoire d'Informatique Fondamentale et Appliquée de Tours (LIFAT), 64, Avenue Jean Portalis 37200 TOURS
prénom.nom@univ-tours.fr
https ://lifat.univ-tours.fr/
***Laboratoire AGORA, 33 Boulevard du port 95011 Cergy-Pontoise
prénom.nom@u-cergy.fr
https ://www.u-cergy.fr/fr/laboratoires/agora.html
****ERTIM-InaLCO, 3 bis rue Taylor 75010 Paris
prénom.nom@inalco.fr
http://www.er-tim.fr

Résumé. Le présent article introduit un thesaurus enrichi relatif aux phénomènes de la nomination et de la référence, construit pour la linguistique, l'analyse de discours et le TAL. Nous détaillons les étapes et les méthodes employées lors de son élaboration, la ressource de type « folksonomie » adossée, ainsi que les expériences d'intégration partielle des connaissances à partir de ressources de connaissance pré-existantes, afin de réduire l'effort humain nécessaire à la construction du thésaurus.

1. Introduction

L'étude de la construction et de la stabilisation du sens en discours, au cœur des recherches en analyse de discours (AD), s'avère également pertinente pour de nombreuses applications en traitement automatique des langues (TAL) et ingénierie linguistique (veille sociale, analyse d'opinion, recherche d'information…). Comme l'ont remarqué (Siblot 1997, 2001), (Frath, 2015),

(Calabrese et Mistaen 2016) et plus récemment (Jackiewicz et Pengam 2018), la notion de nomination, dynamique et contextuelle, permet de renseigner non seulement sur le sens actualisé en discours, mais aussi sur la prise de position des locuteurs sur l'entité nommée. Ainsi, *musulmans modérés, appropriation culturelle*, *réfugiés climatiques*… sont des exemples d'expressions linguistiques relativement instables, rattachées à des questions socialement vives, dont la signification semble se négocier essentiellement en discours.

Considérés dans le cadre discursif, l'acte de nommer et l'usage des nominations se manifestent par un ensemble de traces pouvant être observées dans les corpus. Cependant, le référentiel terminologique relatif à la nomination dont le rôle est de déterminer et de conceptualiser la nature de ces traces semble nécessiter une systématisation et une harmonisation. Les ressources existantes (imprimées ou numériques) souffrent, selon les cas, d'incomplétude, d'obsolescence, de redondance terminologique, de généralité des définitions ou de leur absence (pour des termes dits « orphelins »[1]), ce qui rend difficile leur utilisation pour caractériser les phénomènes repérés dans les textes[2].

Dans le présent article, nous introduisons un thésaurus enrichi constitué dans le but de regrouper et d'harmoniser – au sein d'un système notionnel cohérent et opératoire – les termes issus des travaux en AD, en linguistique et en TAL, relatifs à la nomination et à la référence. Un tel référentiel permet d'éclairer conceptuellement ces deux phénomènes langagiers. Le travail réalisé dans ce cadre permet également d'amorcer une réflexion sur l'appariement entre une sémantique de la nomination ainsi rendue opératoire et les besoins réels en veille (politique, sociétale…).

2. Domaine et finalités du thesaurus

La construction d'un thésaurus dédié plus spécifiquement à l'étude des phénomènes de nomination et de référence a été motivée par le besoin de mieux cerner les rapports complexes entre les entités référentielles et les expressions choisies pour les désigner.

1 Les termes « orphelins » sont des termes introduits dans les textes de spécialité sans définition explicite.

2 Ressources telles que, notamment, *Thesaulangue* et *TermTLF* intégrées dans *Termsciences*, portail terminologique développé par l'INIST en association avec le LORIA et l'ATILF (http://www.termsciences.fr/).

2.1. Les mots et les choses

Le terme « nomination » renvoie à l'acte d'attribution d'un nom à une entité, ainsi qu'au résultat de cet acte. La nomination est une opération linguistique et cognitive, indissociable des processus d'appréhension et de catégorisation des réalités[3]. Elle possède une dimension discursive et dialogique, car l'expression choisie pour nommer un référent reflète la position que le sujet parlant adopte à son égard. Les nominations s'inscrivent enfin dans une dynamique des relations sociales et révèlent des représentations que les locuteurs construisent, négocient et font circuler.

Sur le plan lexical et discursif, la problématique de la nomination touche à la question d'ajustement (adéquation) entre termes ou expressions (dénominations, désignations…) et référents (réalités perçues, vécues, construites…). De nombreux linguistes, dont Authier-Revuz (1995 : 507-520), Culioli (1991), ont étudié la non-coïncidence, le non-un constitutif du rapport de la langue et du monde, en insistant sur l'illusion de la transparence des mots et de l'évidence des choses. Cet écart est d'autant plus sensible que la réalité à verbaliser est complexe ou problématique : instable ou évolutive, émergente, hypothétique ou seulement visée, chargée d'enjeux contradictoires.

Selon les cas, la réponse à ce besoin sera apportée par une innovation lexicale, un emprunt à une autre langue, une néologie de sens, une spécialisation ou une généralisation sémantique. Ainsi, la nomination « patriotisme économique » issue du langage de l'extrême droite française a acquis un sens plus général et une polarité neutre voire positive avec la mise en avant du *made in France* par le ministre du Redressement productif Arnaud Montebourg en 2012.

2.2. Un thésaurus et une méthodologie

Le thésaurus TNR est destiné à être articulé à un modèle linguistique, un schéma d'annotation et un outil d'annotation manuelle.

L'un des objectifs du projet ANR TALAD vise, en effet, à construire une méthodologie générale de repérage et d'analyse des expressions à valeur géné-

3 « La propriété première de la nomination qui, en même temps qu'elle catégorise l'objet nommé, positionne l'instance nommante à l'égard de ce dernier. » (Siblot, Paul, 1997, « Nomination et production de sens : le praxème », *Langages*, n° 31 (127), p.42)

ralisante [4]susceptibles de mettre en évidence des catégories. Ce sont des entités émergentes et relatives, plutôt que des réalités ultimes et absolues, qui sont visées, dans la mesure où le travail d'élaboration est explicitement marqué dans les discours. L'attention est portée tout particulièrement sur la labilité discursive catégorielle des locuteurs. La démarche repose sur l'observation systématique du cotexte de ces expressions afin d'y identifier les traces des différentes formes d'élaboration discursive (intralocutive, interlocutive et interdiscursive) lesquelles fondent et accompagnent l'acte de nomination. Le cotexte est un contexte riche articulant plusieurs catégories de marques linguistiques. Les termes du thésaurus sont destinés à annoter et à caractériser ces marques, suivant un parcours interprétatif défini par la méthode.

3. Thésaurus de la nomination et de la référence (TNR)

3.1. Acquisition des connaissances expertes

Les études de la nomination et de la référence font appel à plusieurs domaines et courants d'analyse linguistique : analyse de discours, sémantique, néologie, lexicologie et terminologie, stylistique.

L'étape initiale de l'acquisition des connaissances linguistiques issues de l'analyse de discours (AD) consiste à étudier l'usage et la définition des termes (lorsque cette définition est fournie) par les spécialistes du domaine. Outre l'étude de l'état de l'art, deux approches ont été employées en ce sens.

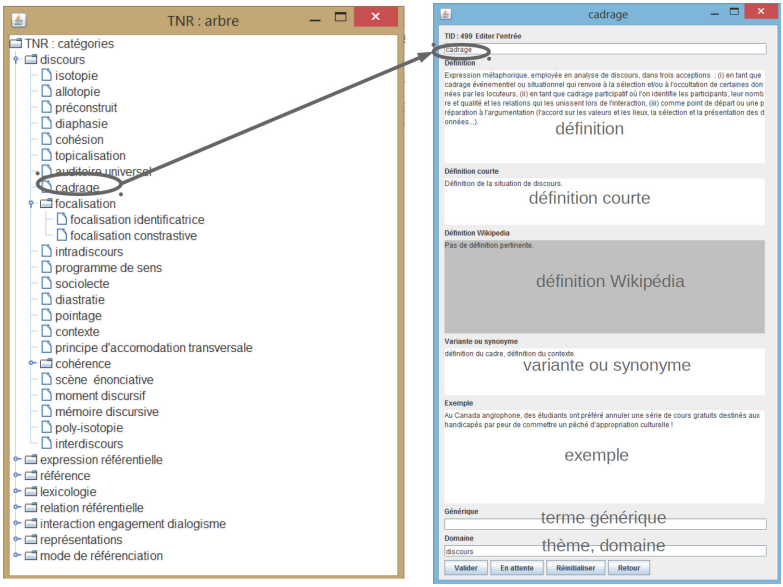
La première approche déployée de façon incrémentale consiste à enquêter auprès des membres de la communauté AD concernés par les problématiques de la référence et de la nomination. Elle permet d'identifier les modèles notionnels et les termes utilisés au sein de la communauté et de faire émerger puis de consolider la structure taxonomique de la ressource. Commencée par de simples listes de termes, cette approche contributive a donné lieu à la création d'un outil dédié qui permet de visualiser et d'enrichir les entrées du thésaurus (figure 1).

À l'heure où nous écrivons, la ressource construite grâce à l'approche experte comporte 372 entrées reparties en 9 catégories générales : *Discours, Expression référentielle, Interaction Engagement Dialogisme, Lexicologie, Mode de référenciation, Référence, Relation référentielle, Représentations,*

4 La *valeur généralisante* est entendue comme une valeur qui permet de rendre général, d'intégrer dans un ensemble d'idée les cas similaires.

Relation de discours. Pour chaque entrée, le TNR contient les informations suivantes : discipline ou aire d'emploi, variante formelle (terme équivalent ou terme approché), statut, définition synthétique, définition étendue, propriétés, relation terme → terme de tête, terme de tête (catégorie, sous ensemble, hyperonyme ou méronyme), terme associé, antonyme (ou terme complémentaire), exemple, catégorie, terme anglais, auteurs et références, publication de référence.

L'interface de complétion par les experts (vue partielle *figure 1*) permet de choisir des champs à afficher (les champs sur lesquels on souhaite travailler). En termes d'expressivité, le thesaurus apparaît comme une ressource riche en informations structurées qui tend vers un dictionnaire collaboratif. La présence des termes vedettes et variantes, des hyperonymes, des termes similaires permet son exploitation ultérieure dans le cadre d'analyse sémantique des textes de spécialité.



Les liens vers les autres ressources (notamment, thesaurus bilingues) sont en train d'être construits sachant que les ensembles des termes communs à ces ressources externes et au TNR sont de taille assez réduite et qu'il s'agit principalement des ressources de terminologie linguistique généralistes.

À ce jour, nous avons exploré les ressources pré-existantes suivantes :

- les ressources répertoriées sur le portail Termsciences (ressources terminologiques génériques) telles que Lexique383, OpenLexicon, Thesaulangue ;
- « Guide terminologique pour l'analyse des discours » (de Nuchèze et Colette, 2002), ressource livresque ;
- *SIL Glossary of Linguistic Terms*[5], glossaire généraliste de termes linguistiques bilingue français – anglais. Ce glossaire contient 8 600 entrées pour le français dont assez peu concernent l'analyse de discours. Pour le terme « discours » :
- nous avons repéré 14 termes composés ayant le terme discours comme tête syntaxique : « discours + *expansion* », *expansion* ∈ *{argumentatif, authentique, cité, d'exhortation, d'exposition, d'instructions, descriptif, dialogique, direct, indirect, indirect libre, monologique, narratif, rapporté}*.
- Parmi les termes ayant « discours » comme expansion « *tête*+(Préposition)?+ discours », l'on trouve *tête* ∈ {*type, genre, grammaire, analyse*}. Par ailleurs ce glossaire inclut les termes *interdiscours*, *métadiscours*.

L'amorçage et la construction collaborative du thesaurus ont révélé que la production terminologique importante est caractéristique des pratiques de la communauté d'analyse de discours. Cette observation a induit la démarche semi-automatique d'extraction des termes candidats à partir des textes de spécialité afin de se donner les moyens d'observer l'activité terminologique des linguistes et des notions qui sont « en chantier » au sein de cette communauté. L'abondance des termes, la spécification des termes existants témoignent de la spécialisation du domaine d'analyse de discours. Cependant, cette abondance peut également compter des cas de redondance inutile.

Outre la construction par les experts du domaine, l'approche semi-automatique a été expérimentée sur un corpus de 40 articles scientifiques traitant de la nomination et de la référence, un sous-ensemble de corpus constitué pour l'état de l'art. Ce sous-corpus de 1 115 115 mots a permis de procéder à une extraction terminologique avec l'outil *TermSuite* (Rocheteau et Daille, 2011) afin d'obtenir des termes-candidats. Cette extraction a eu pour objectif :

- la preuve d'adéquation (notamment, en termes de couverture) du thesaurus constitué par les experts dans une démarche descendante ;

5 https://feglossary.sil.org/

- l'identification des termes-candidats qui pourraient enrichir le thesaurus.

Dans le cadre de cette approche alternative, de nombreux candidats ont pu être extraits automatiquement. Après une pré-validation semi-automatique (par règles définies manuellement qui concernent l'inclusion lexicale et la structure des termes), un ensemble de candidats pré-validés (**#pré-validés**) qui correspond à un pourcentage du nombre des candidats issus de l'extraction automatique (**%pré-validés**) a été obtenu. Parmi ces termes pré-validés 91 % (**validés**) ont pu être retenus pour étude en vue de leur intégration dans le thesaurus (Tableau 1).

extraits	**# pré-validés**	**% pré-validés**	**#validés**	**%validés**
13 885	542	3 %	493	91 %

Tableau 1. Acquisition semi-automatique des termes candidats

Parmi les termes extraits validés, on trouve par exemple *applicabilité référentielle, allocutaire, propriété événementielle, saillance du cadrage* etc.

Dans la suite de l'expérience d'acquisition automatique, nous avons tenté de caractériser ces termes-candidats du point de vue de leur distribution, les rapprocher des termes déjà contenus dans le thesaurus et des autres termes candidats. Pour cela, nous avons considéré une méthode distributionnelle fondée sur le calcul des plongements des mots (*word embeddings* (Mikolov, 2013)). Les mesures *cosinus*[6] calculées pour explorer la similarité existant entre les vecteurs des mots obtenus à partir du sous-corpus qui a servi pour l'extraction terminologique ont permis de faire les rapprochements détaillés dans le tableau 2.

6 $\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$ où A et B sont les vecteurs obtenus par plongement lexical. Il n s'agit pas de cosinus d'angle, les valeurs négatives (mots opposés) ont été peu considérés dans le cadre de la présente expérience.

Terme 1	Terme 2 (mesure cosinus)	**Qualification et remarques**
applicabilité référentielle	charge dialogique (0,84) inapplicabilité référentielle (0,88) statuer (0,95), allocutaire (0,88), congruence (0,87), degré (0,85), enchaînement (0,80)	Rapprochement d'un terme candidat et terme déjà dans le TNR. Proposition d'un terme au sens opposé. Description du terme candidat.
formant signalétique	description définie (0,82) d'agent (0,86), différenciateur (0,86) adverbe (0,85)	Proposition du terme associé (termes 1 et 2 déjà dans le thésaurus). Spécification du terme (pistes)
polarité	déviance (0,88), idéalités (0,87), spécifiante (0,81), morale (0,84) Gosselin (0,87)	Introduction du terme candidat (termes similaires). Suggestion d'une référence bibliographique
onomastique (nom)	sémantique (0,92), mono-référentielle (0.87), formants (0,83), d'entités (0,82)	Introduction du terme candidat.
rhétorique éristique	dénotation (0,88), étiquetage (0,86), intension-extension (0,85)	Termes associés à un terme issu du TNR.
étiquetage accusateur	diabolisant (0,97), fasciste (0,91), nazisme (0,89), supprimer (0,86), régime (0,86), abus (0,84), dictature (0,82), ethnocide (0,82), meurtre (0,81), victimes (0.80)	Exemplification du terme déjà présent dans le thésaurus.

Tableau 2. Exemples d'informations obtenues à partir de l'analyse de similarité distributionnelle.

Par ailleurs, la distribution des termes permet d'étudier les descripteurs déjà présents ou pouvant être intégrés dans le TNR : les plongements des adjectifs « catégoriel », « sémantico-syntaxique », « sémantico-référentiel » sont assez proches dans l'espace vectoriel considéré dans le cadre de l'expérience.

Même si la taille du corpus des articles scientifiques est insuffisante pour permettre des résultats fiables strictement issus des méthodes d'apprentissage automatique et des approches neuronales, cette première étude combinant l'extraction et la pré-caractérisation automatique semble permettre la réduc-

tion de charge de travail des experts humains dans le cadre de la construction d'une ressource de spécialité. Son application est envisagée pour le traitement de corpus plus vastes d'articles scientifiques en support à la construction experte.

L'intersection entre les résultats obtenus issus des méthodes semi-automatiques est de 83 termes (repérés automatiquement et déjà présents dans le thesaurus). Cette intersection a été exclue de l'ensemble des termes-candidats. Une expérience sur un corpus plus vaste serait nécessaire pour se rendre compte de la couverture du thesaurus en cours de construction.

Le TNR est destiné à guider la mise en place d'un environnement qui l'inclut comme ressource terminologique, mais dispose également d'un modèle linguistique, d'un schéma d'annotation et d'un outil d'annotation par les humains ; des ressources complémentaires de type « folksonomie » pour stocker la connaissance non terminologique qui permet la pré-annotation automatique des textes.

4. Folksonomie TNR : structuration et exploitation

4.1. Motivation

En termes d'expressivité, le type de ressource recherché est celui d'une ressource notionnelle riche et originale, qui n'est spécifiquement ni un dictionnaire, ni un thésaurus, ni une ontologie, mais – en puissance – un peu tout cela. Les niveaux de structuration de la connaissance sur la référence et la nomination (rendre la connaissance opératoire) et de représentation (permettre des sorties en utilisant de différents formats dont ceux du Web sémantique) ont été séparés.

Le but de cette démarche est de permettre un maximum de souplesse quant à l'acquisition et à la structuration des connaissances pertinentes pour la tâche visée, afin de pouvoir répertorier et caractériser finement les traces des phénomènes linguistiques et discursifs à l'œuvre dans les processus de référence et de nomination.

4.2. Mise en œuvre

Pour la systématisation des traces des opérations concernées (définition, prise en charge, relations sémantiques…), nous avons construit une base de connaissances (« folksonomie ») sous forme de graphe. Les nœuds de ce

graphe sont des termes du thésaurus et les items lexicaux (patrons lexicaux et méta-discursifs, segments textuels pertinents pour l'étude de la nomination, items lexicaux). Les relations sont

- des relations faiblement typées entre les nœuds représentant les termes du thesaurus et les nœuds représentant les items lexicaux (traces pertinentes pour l'étude de la nomination et de a référence repérés dans les textes) ;
- des relations sémantiques et discursives associées aux items lexicaux.
- Pour réduire le coût de l'acquisition des termes et relations des items lexicaux, nous avons exploré l'utilisation des ressources telles que :
- réseau lexico-sémantique de connaissance générale pour le français RezoJDM (Lafourcade 2007)[7] ;
- ASFALDA[8], FrameNet (Ruppenhofer *et al*. 2016) pour le français ;
- corpus annoté ANNODIS[9] (exploitation des relations discursives actuellement à l'étude).

La folksonomie$_{TNR}$ est parcourue dans le cadre de test d'annotation automatique afin de détecter les traces potentielles de construction de sens des nominations émergentes à partir des traces connues et déjà répertoriées et rattacher ces traces potentielles aux entrées du thesaurus. La structure de données obtenue est celle d'un graphe avec différents items pertinents pour l'annotation en termes des phénomènes de nomination et de référence : *terme_source, type_de_relation, poids, origine, terme_cible*. Les exemples de relations obtenues sont donnés ci-dessous.

(1) information d'ordre taxonomique et variantes

acronyme, r_definition, 111, tnr, Mot formé des initiales ou éléments initiaux de plusieurs mots, prononcé comme un mot ordinaire.
acronyme, r_example, 111, tnr, SNCF
acronyme, r_isa, 111, tnr, expression référentielle
acronyme, r_isa, 148, jdm, sigle
pseudonyme, r_isa, 111, tnr, anthroponyme
pseudonyme, r_syn, 84, jdm, faux nom

(2) termes polysémiques (relation de raffinement r_raff _sem explicite les sens possibles)

7 http://www.jeuxdemots.org
8 http://asfalda.linguist.univ-paris-diderot.fr/frameIndex.xml
9 http://redac.univ-tlse2.fr/corpus/annodis/

type d'entité, r_domain, 111, tnr, TAL (traitement automatique du langage)
type d'entité, r_raff_sem, 43, jdm, type d'entité >informatique
type d'entité, r_raff_sem, 45, jdm, type d'entité>caractéristique

(3) relations sémantiques

pronom personnel, r_lieu, 98, jdm, phrase *(relation lieu typique)*
point de vue, r_carac, 30, jdm, partagé>commun *(relation caractéristique typique)*
point de vue, r_holo, 17, jdm, débat d'idées *(relation partie-tout)*

(4) lien vers les cadres («frames») correspondants

admettre, r_frame, 50, framenet, FR_Agree_or_refuse_to_act
admettre, r_frame, 50, framenet, FR_Awareness-Certainty-Opinion
admettre, r_frame, 50, framenet, FR_Being_in_favor_of
admettre, r_frame, 50, framenet, FR_Statement-manner-noise

Ces exemples montrent que la représentation sous forme de graphe permet d'harmoniser les informations issues des différentes ressources et processus. Le poids des relations est actuellement fixé par défaut, il fera l'objet d'une harmonisation ultérieurement. La folksonomie$_{TNR}$ contient 12 515 relations dont 2 048 relations sont issues du TNR (restructurent le TNR sous forme relationnelle), 8 101 - du RezoJDM à ce jour, 1 574 - du FrameNet français, 792 - des contributions et listes (il s'agit, en particulier, des patrons méta-discursifs).

4.3. Exploitation

L'exploitation de la folksonomie$_{TNR}$ concerne l'analyse et la pré-annotation des textes. Ces processus sont en train d'être construits à l'heure où nous écrivons. La pré-annotation exploite les patrons méta-discursifs et les cadres (*frames*).

Exemple 1 : **représentation d'un patron méta-discursif** (segment textuel «réflexion à mener en France sur cet islam modéré»).

Patron méta-discursif : [$X :Nom r_isa acte de penser] en [$Y :Nom r_isa pays] sur [adj :dem] $Z :Nom
Patron (forme relationnelle) :
acte de penser, *r_pattern*, [$X :Nom r_isa acte de penser] en [$Y :Nom r_isa lieu] sur [adj :dem] $Z :Nom
(dans la folksonomie$_{TNR,}$ le patron apparaît comme une chaîne de caractères, poids de la relation r_pattern est fixé par défaut)

Exemple 2 : **pré-annotation automatique du segment** «on observe avec dédain la prolifération de cette pratique» :

patron : quand on observe $X :Nom :	**cadre** : Judgment	**relations** : regarder—r_instr-->mépris regarder--r_instr-->dédain

À titre exploratoire et grâce à un ensemble des patrons lexico-sémantiques et méta-discursifs classés par catégorie pour l'expérience, il a été possible de quantifier les éléments présents dans un corpus (66 359 mots) constitué dans l'objectif d'étudier l'élaboration de la nomination « musulmans modérés », corpus décrit dans (Pengam et Jackiewicz, 2019).

catégorie	**#occurrences**	**%occurrences**
Signalement minimal (présence des guillemets)	29	18,2 %
Signalement ou avertissement	35	22 %
Désignation	7	4 %
Élaboration (définition ou explicitation)	5	3 6%
Reprise, citation, adhésion, interaction	6	3,6 %
Distanciation critique et reformulation	40	25 %
Cadre de validité	16	10 %
Rejet et renomination polémique	21	13,2 %
Total	159	100 %

Tableau 3. Pré-annotation

La pré-annotation automatique grâce à une ressource de connaissance riche est un outil puissant de structuration notamment en ce qui concerne l'élaboration des modèles linguistiques et des schémas d'annotation. Les intuitions qui ont émergé de l'expérience de pré-annotation ont appuyé la mise en place d'une campagne d'annotation par les annotateurs humains. Cette campagne a été axée sur le phénomène de nomination émergente suscitant ou non la controverse (*musulman modéré, appropriation culturelle, mobilité douce, écofascisme*). 230 segments textuels (co-textes) de nomination émergente ont été annotés en termes d'analyse de discours. Le schéma d'annotation a intégré 3 aspects : le plan ontologique ou linguistique (par exemple, controverse sur le phénomène ou sur le terme utilisé pour le nommer), le procédé (introduction, ajustement ou rejet) et l'attitude (prise en charge, interaction, cadrage). Outre ces angles d'analyse, les relations sémantiques (en particulier, les relations statiques au sens de Desclès (2013)) ont été intégrées dans le schéma d'annotation.

5. Conclusion : vers une ontologie de la nomination et de la référence

Nous avons décrit les différentes expériences qui ont été menées dans le but de structurer, enrichir et harmoniser les connaissances théoriques liées à l'étude de la nomination et de la référence, puis de proposer une ressource de connaissance intermédiaire permettant de lier les termes qui servent à nommer les phénomènes à la réalisation de ces phénomènes dans les textes.

Les perspectives de ces travaux concernent la stabilisation des ressources que nous avons décrites ainsi que l'interopérabilité à la fois du TNR appelé à évoluer vers une ressource termino-ontologique et des données structurées issues des campagnes d'annotation que nous avons évoquées. L'effort d'intégrer les ressources de connaissance pré-existantes témoigne de la volonté d'aboutir à une ressource dotée d'une interopérabilité de contenu. L'interopérabilité de format du TNR peut être obtenue grâce à l'utilisation des formats interopérables tels que :

- TBX[10] (glossaire dans un format d'échange des terminologies, ISO 30042) ;
- Lemon OntoLex[11] (modèle de représentation dans le format du web sémantique qui étend le format OWL afin de permettre de capturer les informations lexicales et linguistiques associées à des concepts d'une ontologie) ;
- SKOS[12] (format d'organisation des connaissances de type thésaurus) .

Celle des données de sortie des campagnes d'annotation est atteinte grâce au développement d'une ressource dédiée à l'annotation qui permet le format de sortie interopérable (brat, format compatible avec l'outil d'annotation brat). L'outil permet d'ajouter d'autres formats de sortie interopérables.

Bibliographie

Calabrese, Laura et Valériane Mistaen, « La nomination des migrants dans Le Monde et Le Figaro. Analyse d'une catégorisation polémique », REFSICOM [en ligne], Médias et migrations/immigrations 1. Des

10 https ://www.iso.org/standard/45797.html
11 https ://www.w3.org/community/ontolex/wiki/Main_Page
12 https ://www.w3.org/2004/02/skos/

représentations aux traitements des médias traditionnels, mis en ligne le 23 novembre 2018, consulté le mercredi 04 décembre 2019.

Cance, Caroline, et Danièle Dubois. « Dire notre expérience du sonore : nomination et référenciation », *Langue française*, vol. 188, no. 4, 2015, p. 15-32.

Cislaru, Georgeta (dir.) *et al*. *L'acte de nommer : Une dynamique entre langue et discours*. Nouvelle édition [en ligne]. Paris : Presses Sorbonne Nouvelle, 2007 (généré le 8 février 2019).

Desclés Jean-Pierre. Interactions entre langage, perception et action. In : *Faits de langues*, n° 1, mars 1993. Motivation et iconicité. p. 123-127.

Fellbaum, Christiane, 1998, WordNet : An Electronic Lexical Database. Cambridge, MA : MIT Press.

Frath, Pierre, 2015, « Dénomination référentielle, désignation, nomination », Langue française, n° 188, p. 33-46.

Firas Hmida. Identification et exploitation de contextes riches en connaissances pour l'aide à la traduction terminologique. Informatique et langage [cs.CL]. Université de Nantes, 2017. Français.

Hoffart, Johannes, Suchanek, Fabian M., Berberich, Klaus, and Weikum, Gerhard. 2013. YAGO2 : A spatially and temporally enhanced knowledge base from Wikipedia. Artif. Intell. 194 (January 2013), 28-61.

Jackewicz, Agata, Pengam, Manon, 2018, « Des *musulmans modérés* dans les discours médiatiques. Etude linguistique d'une expression controversée », *Colloque international Les représentations médiatiques de l'islam et des musulman.e.s,* 19-20 juin 2018, Versailles Saint-Quentin-en-Yvelines, France.

Lafourcade, Mathieu, 2007, *Making people play for Lexical Acquisition*. In Proc. SNLP 2007, 7th Symposium on Natural Language Processing. Pattaya, Thailande, 13-15 December 2007, 8 p.

Longhi Julien, « Stabilité et instabilité dans la production du sens : la nomination en discours », Langue française, 2015/4 (N° 188), p. 5-14.

Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Greg, and Dean, Jeffrey. 2013. Distributed representations of words and phrases and their compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'13), C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.), Vol. 2. Curran Associates Inc., USA, 3111-3119.

Rocheteau, Jérôme et Daille, Béatrice, 2011. TTC TermSuite : A UIMA Application for Multilingual Terminology Extraction from Comparable Corpora. Proceedings of the 5th International Joint Conference on Natural Language Processing.

Ruppenhofer, Josef ; Ellsworth, Michael ; Petruck, Miriam R. L. ; Johnson, Christopher R. ; Baker, Collin F. ; Scheffczyk, Jan, 2016, FrameNet II : Extended Theory and Practice (revised ed.). Berkeley, CA : International Computer Science Institute.

Siblot, Paul, 1997, « Nomination et production de sens : le praxème », *Langages*, n° 31 (127), p. 38-55.

Abstract

The present article introduces a rich thesaurus built for the discourse analysis domain and focused on nomination and reference issues. We detail the stages and the methods that have been used in the framework of the thesaurus building process. Our experiences of structuring the thesaurus as well as a « folksonomy » supporting it aim at building an ontology for nomination and reference. Such ontology will be designed for natural langage annotation in terms of nomination phenomena. Thus, human annotation campains and automatic annotation tests accompany our experiences of knowldge structuring and representation. We also explored different possibilities in order to reduce human effort necessary for designing a specialized resource.

Towards a Model for Creating an English-Chinese Termbase in Civil Aviation

Hui Liu*, Xiao Liu**

*P. O. Box 1012
College of Foreign Languages
Nanjing University of Aeronautics and Astronautics
No. 29, Jiangjun Road, Jiangning District
211106 Nanjing, Jiangsu Province, P.R.C.
luisaliu0339@gmail.com
** P. O. Box 1012
College of Foreign Languages
Nanjing University of Aeronautics and Astronautics
No. 29, Jiangjun Road, Jiangning District
211106 Nanjing, Jiangsu Province, P.R.C.
maggieliuxiao@nuaa.edu.cn

Abstract. At present, precise communication in a special field is impossible without appropriate terminology. It is even so for civil aviation which is a fast-developing, interdisciplinary field and is in need of extreme precision. Based on a linguistic analysis of terms in the field of civil aviation (CA), this paper explores the key aspects related to the construction and compilation of a bilingual termbase (terminological database) in this field and aims at highlighting the imperative need for such a tool to support the effective exchange of knowledge and information in a professional setting. The design principles of this termbase cover three key aspects : the selection of terms and pictures, display options as well as division of roles and functions. The practice and concepts may also be extended to create termbases for other languages and fields.

1. Introduction

As the key component of domain-specific communication, "terminology plays a crucial role wherever and whenever domain-specific information and knowledge is generated, used, recorded and processed, passed on, imple-

mented, translated and interpreted (Infoterm 2005)." As a result, the development of terminologies has undeniably become a top priority in order to meet the requirements of the 21st century specialized communication, let alone "the various uses of civil aviation data such as analyses related to safety, security and the efficiency of civil aviation and its environmental impact as well as forecasting and planning require a suitable classification and a clear definition of civil aviation activities (ICAO 2009)". At present, there are various mono-, bi- and multilingual terminological dictionaries and glossaries in the forms of hardcopy, electronic, and online resources in major fields available on the global market. Besides, with the help of software technologies and a series of national as well as international terminology programs and management entities, large multilingual termbases covering hundreds of thousands of, or even millions of terms, in specialized domains, have become a reality (see, e.g. IATE- EU's multilingual termbase, TERMIUM Plus- Canadian government's terminology and linguistic data bank, Termonline-the bilingual termbase of China National Committee for Terms in Sciences and Technologies, etc.).

However, although it is essential to have authoritative resources in the field of civil aviation (CA), the number of dictionaries, termbases and terminological products lags far behind than that in other fields. The reasons are multifold. Firstly, none of the above-mentioned authoritative termbases deal with this particular domain. Although there are some online tools that support the search and management of terms in CA, they are generally monolingual or mono-functional. Moreover, with rapidly expanding air traffic worldwide, CA is more and more related to various fields including chemistry, physics, geography, astronomy, computer science, and communications. Terminology in CA is also influenced by the concepts and terms in the above fields. As a result, the harmonization and standardization of terminology in the field of CA in the context of communication requirements in both Chinese and English is imperative.

2. Previous Researches and State-of-the-Art

2.1. CA Dictionaries, Termbases and Other Resources

To begin with, English CA terminology problems are often tackled (and hopefully solved) with a dictionary, e.g. *An Illustrated Dictionary of Aviation* by B. Kumar *et al*. (2002), *Dictionary of Civil Aviation* by R.K.C. Shekhar & S.K.C. Shekhar (2005), *Dictionary of Aeronautical Terms* by D. Crane

(2012) and *The Aviation Dictionary for Pilots and Aviation Maintenance Technicians* by D. Jones (2003). Of these dictionaries mentioned, the latter is more comprehensive and targeted primarily at, as its title suggests, pilots and maintenance technicians. In addition, there are some online glossaries and dictionaries that include terms belonging to traditional fields of aviation, or more restrictively, subfields of CA, say, for instance, air transport and traffic control. If one conducts a Google search for the phrase *civil aviation dictionary*, it yielded 6,120,000 results (as on May 6th, 2019). However, a similar research for the number of dictionaries in anther field, say, business, yielded 364,000,000 results. As a result, the number of results for *civil aviation dictionary* is less than 2 percent of that for *business dictionary*.

So far as the situation concerning the Chinese CA dictionaries, it is even worse. This may be partially due to the fact that in China CA is still a very young field of expertise, which gained popularity and significance only after the country aligned with ICAO (International Civil Aviation Organization) and the subsequent adherence to related regulations in 1983. There are only a few dictionaries available (e.g. D. He & R. Zhou 1997 ; S. Zhu 2013 ; T. Hu *et al*. 2016) and their primary focus is on a related field instead of CA, namely, aeronautics or astronautics. For example : *The English-Chinese Dictionary of Aeronautical and Astronautical Engineering* by B. Liang (2000) and *An English-Chinese Dictionary of Aeronautic Technology Abbreviations and Acronyms* by M. Huang (2015).

Secondly, despite of the fact that there are some internationally-acknowledged multilingual termbases which include millions of terms, none of them deals with this particular domain. Although there are some online tools that support the search and management of terms in CA (e.g. Termbox[1] and Yunfan Online Dictionary[2]), they are generally monolingual or provides either the management or search function.

Meanwhile, symbols, markings and signs are equally important in CA, e.g. in safety management of passenger cabins (F. Zhang 2017) as well as aerodrome design and operations (ICAO 2018). However, they are absent from most of the resources mentioned above. These non-verbal forms should also be included in a terminological product.

1 An online term management solution provider based in China, see http://termbox.lingosail.com/.

2 An online termbase in CA with bilingual search function only, see http://www.cadict.net/.

Moreover, since the establishment of China's own CA industry, although the subordinate units and organizations have compiled some glossaries in the form of pamphlets and distributed them on a small scale, they have not yet officially gained a comprehensive coverage and authoritativeness. Not only is the *Terms in Aeronautical Science and Technology* (2004) issued by China National Committee for Terms in Sciences and Technologies[3] difficult to include all the terms in CA, but it is also difficult to find references for terms in CA on a wider Chinese-speaking are (e.g. in Taiwan). Only a few of the airlines have established their own termbases with a small number of term entries and these termbases can only be used on the intranet.

In summation, the absence of a comprehensive terminological product in the field of CA urgently requires an authoritative and multifunctional terminology management system and product. As mentioned earlier, all professions need a clear and consistent formulation of field-specific terminology without which a professional communication and practice simply cannot exist. And that is why the creation of a proper bilingual CA termbase based on a scientific management of terminological data according to logical structure of knowledge is of vital importance.

2.2. Features of Terms in the Field of CA

An analysis concerning the features of terms in the field of CA is provided in this section. English and Chinese examples are given in small capitals and italics respectively, with a symbol "><" between them to indicate the direction of translation (from English into Chinese or vice versa).

2.2.1. Semantic Analysis of CA Terms

Homonymous terms abound in the field of CA, which is in accordance with Cabré's statement that terms are likely to develop homonymous relations rather than polysemous ones (1999, 108-111). For instance, the term *abandon* has two meanings, 1) to bail out or eject out of an aircraft and let it crash and 2) to talk away or leave an aircraft on the ground in an emergency as when it is on fire (B. Kumar *et al*. 2002, 11). And the term *aberration* has three meanings: 1) a condition in an optical system in which the images are imperfect

3 China National Committee for Terms in Sciences and Technologies is an authoritative organization authorized by the State Council of China to examine and publish scientific and technological terms on behalf of the country.

or improperly located, 2) geometrical inaccuracies introduced by optical, IR (infrared), or similar electromagnetic systems in which radiation is processed by mirrors and 3) the displacement of the apparent directions of the stars resulting from the motion of the observer (*ibid*.).

In addition, synonymy, i.e. the phenomenon of more than one term to designate a single concept, seems also to be present to a great extent in CA terms, and in both English and Chinese languages. For instance, *设计者/设计师*>designer and destruction test/rapture test>*破坏试验*.

2.2.2. Morphosyntactic Analysis of CA Terms

Nouns and verbs predominate in CA terms and most of both nouns and verbs are polylexcial units. In addition, there are a small number of terms made up of a single word, which are either simple terms or complex terms formed by a series of means, e.g. compounding, affixation, clipping or conversion.

The most frequent affixes for terms in the field of CA are agentive suffixes -er and gerund suffixes -ing. Due to the pragmatic requirements of terminology, i.e. the need for concision of a terminological unit (M. Milić 2015), there are also a great many clippings, of which the majority are acronyms and initialisms. Acronyms include names of related organizations and authorities, e.g. ICAO (International Civil Aviation Organization), IATA (International Air Transport Association) and FAA (Federal Aviation Administration). Initialisms are also abundant. For instance, BUEC for backup emergency communications and SFCIN for specific fuel consumption installed.

Polylexical units, or sometimes called as multi-word terms (Babić 1990, 36), phrasal lexemes (L. Lipka 2002, 79), or terminological phrases (S. Ledrew 1997, 31), differ from other terms in that they consist of more than one word and may include collocations (e.g. access taxiway>*进入滑行道*), syntagms (e.g. active runway>*现用跑道*), clauses (e.g. request further climb>*请求进一步上升*) as well as full sentences (e.g. Liquids may only be carried within separate containers, each of which with a volume no greater than 100ml. >*液体物品只可装入独立容器内，每个容器容积不超过100毫升。*), especially when it comes to official signals. Due to the language differences, i.e. Chinese is a non-inflectional language while English an inflectional one, these terms, or polylexical units, share the similarity merely in the nature of polylexicality, as the constituent words are governed by language-specific morphosyntax.

2.2.3. Features of Chinese Terms

Wüster proposes that "terms should be treated differently from general language words (Pearson 1998, 10)." Six features of Chinese Terms in the field of CA are acknowledged by the academic circle (see Q. Zhou 2010, 41-46 ; X. Liu & J. Zhou 2011, 174 ; J. Wen & P. Li 2011, 27-32 ; D. Huang 2017, 108-114). Firstly, Chinese Terms in the field of CA are more often than not monosemous. In other words, one term designates one concept, or vice versa. Therefore, a one-to-one correspondence between a term and a concept is usual.

Secondly, in accordance with H. Guo that "an expression, if it rises to the philosophical and theoretical level from the perspective of social life and experience, then the arbitrariness and flexibility of its expression will be lost or at least greatly reduced (2006, 102) ", Chinese Terms in the field of CA are quite stable.

Thirdly, abbreviations are abundant in CA terms in Chinese. This is partly due to the fact that in order to enable the terms to fully reflect the characteristics of the concepts, term users and creators often resort to phrases prior to abbreviations. For example, before laser (the abbreviated form) is formally recognized as a term, the full name "light amplification by stimulated emission of radiation" is used. *镭射*, the Chinese equivalent of laser, also appeared after *激射光辐射放大* or *光量子放大* (the Chinese counterpart for light amplification by stimulated emission of radiation).

Furthermore, Chinese Terms in the field of CA gradually aligned with *ICAO,* or, tried to follow the same international norms since China's adherence to related regulations in 1983. Since then, Chinese Terms in the field of CA began to show its "internationalism". For example, strategic flow management, pre-tactical flow management and tactical flow management used to be translated into *战略流量管理*, *预战术流量管理* and*战术流量管理*respectively, which left the readers an impression that these words were more of a military nature (because the Chinese character *战*refers to a war). But at present, they are translated into *先期流量管理* (the analysis and prediction of flight flow over a longer period of time, usually more than seven days, so as to confirm the potential flight congestion and propose solutions), *飞行前流量管理* (the analysis and prediction of flight flows over a shorter period of time, usually between one to six days, so as to confirm the potential flight congestion and propose solutions) and *实时流量管理* (on the day of the flight, taking corresponding measures to control the aircraft for an orderly operation

according to the air traffic situation) respectively, which are also more appropriate.

Fifthly, zero-translation is applied when CA terms are to be translated from English into Chinese, because term translation is, by nature, the search for "equivalent terms that represent the same concept" among heterogeneous foreign languages (Z. Feng 1997, 7). Zero translation refers to the use of original words in the target language. Due to the international nature of CA industry, there are a large number of these words, mostly in English, that can be understood by professional readers in the industry whatever their mother tongue is. This translation method is not a temporary or an expedient measure, neither is it limited to the use of internationally accepted abbreviations and symbols. It also includes various instrument display options, signs on doors, lights, warning information and so on.

Finally, with expanding air traffic worldwide, CA is more and more related to various fields including chemistry, physics, geography, astronomy, computer science, and communications. As a result, besides being influenced by English, Chinese terms in CA are also influenced by the concepts and terms in the above-mentioned fields. And terms in this field displays somehow a "cross-domain" nature.

3. The Selection of CA terms

Terms to be included in this specialized CA termbase must be selected carefully, so that the termbase covers all the important terms that are unique to CA or a subfield. Such a termbase should not only include terms, but also symbols, signs and pictures, as they are equally important in CA.

3.1. The Selection Process for CA Terms

Terms in CA can be in the form of single words, phrases or fixed expressions. And there are at least three sources to obtain possible terms in CA, i.e. CA dictionaries, books and manuals as well as suggestions from field and language experts.

Besides English dictionaries like *Dictionary of Civil Aviation* (R.K.C. Shekhar & S.K.C. Shekhar, 2005), English-Chinese CA dictionaries are good sources, if not the best, for term selections in this field. China Civil Aviation Publishing House, a central professional publishing house under the Civil Aviation Administration of China, has published a series of such dictionar-

ies (see, e.g. D. He & R. Zhou 1997; H. Zhang 2011; S. Zhu 2013; T. Hu *et al*. 2016). Among these dictionaries, *An English-Chinese Dictionary of Civil Aviation* (H. Zhang 2011) is by far the most comprehensive English-Chinese CA dictionary. The completion of it took ten years. Including around 30,000 words, this dictionary covers the vocabulary of various subfields of CA. As a result, these words could be used as the basis for the termbase, if permission to use the electronic copies of these dictionaries can be obtained. On the other hand, however, the coverage of CA symbols of dictionaries is quite limited, which is a shortcoming for sources of this kind.

The second source of CA terms is books and manuals such as *An Introduction to Civil Aviation* (D. Liu 2005), *English for Civil Aviation Service* (T. Yu 2011) and *Civil Aviation English* (J. Pu & A. Chen 2012), to name but a few. If permission for an electronic copy of these books or the termlists attached is obtained, these books can be converted into plain text formats and included in the termbase. Some manuals may include indices that carry the most relevant or important words in the field of CA, which is also a good source for the current termbase.

After making a potential termlist based on dictionaries, manuals and books, field and language experts can be consulted. We consider the consultation with experts in a specialized field highly necessary for an LSP as well as a particular termbase because terminologists may not know a specialized subject field so well as field experts do. The field experts whom we consider knowledgeable in the field of CA are lecturers and professors delivering courses related to CA, such as those who teach airline operation and management for degree programs in the university in which the authors are working. In addition to the check of terms collected, experts would be expected to provide suggestions on new terms and a systematic classification of the key terms for the purpose of ensuring that all the necessary terms are included in the termbase. Important symbols, markings and signs of the above mentioned sources can also be included in a termbase with permission.

3.2. The Selection Process for CA Symbols, Markings and Signs

As mentioned before, symbols, markings and signs are integral parts of CA, without which might result in failures or even serious accidents. In addition, a CA termbase should contain more than words because terms can also exist in a non-verbal form. In addition to the words that are selected as head-

words, symbols, markings and signs in the termbase will sometimes be more attractive for the target audience of the terms.

Besides the sources of terms mentioned in the previous section, symbols, markings and signs can be taken from aircrafts, airports, or, when unique symbols are found in aircraft manufacturers. These symbols, markings and signs will be arranged thematically in the CA termbase. Users of the termbase will be able to either see them as a group or search through an individual picture or symbol based on keywords of the description. For instance, a search of the keyword "helicopter" will result in two links-one to the termbase entry for helicopter, and the other to the related sign such as the one in Figure 1 below.

Fig. 1 – The symbol of the helicopter (ISO 2007)

Due to the fact that terms in the termbase are adopted from various sources mentioned above, term selection is completed based on manual work to ensure the correctness. Up till now, over 7,000 representative terms have been included in the termbase with around 100 pictures, symbols and signs.

4. Data Fields and Functions of the Termbase

Among the data categories in *ISO 12620 Management of Terminology Resources - Data Category Specifications* (2019), the following would best serve the termbase and individual needs of users : entry term (both Chinese and English), definition, variant(s), reference and date of record.

Figure 2 is a sample of a default search result for the term "altitude hold" in the termbase. The entry term is provided in both Simplified Chinese and English with definitions and references in both languages.

Fig. 2 – The search result for "altitude hold" in the termbase

We consider entry term as the most important data category in the termbase because "it designates (i.e. names) the concepts in the field (A. Kilyeni *et al.* 2012)." Each term should be recorded based on terminological conventions of presentation, i.e. in the singular, in the infinitive, in the lowercase, and so on, unless the plural form or the capital letter has a terminological significance (*ibid.*). A definition that gives the semantic characteristics which distinguish one concept from all others (S. Pavel & D. Nolet 2001, xix) should also be provided to offer more assistance to users.

In addition to the search function, the advanced search function of the termbase provides different types of matching (e.g. exact match, any word or all words) and the possibility of searches in a special sub-domain (will be discussed in section 5). The search of "symbols, markings and signs" function enables users to search for related elements with a click. For the purpose of providing a dynamic and interactive termbase, suggestions and feedbacks from users are highly welcome. Users can click on this "suggestions and feedbacks" button and fill in the necessary information (e.g. term, definition, context, as well as name and email address) in the pop-up window. Users can also send messages to the administrator of the termbase by clicking on the "contact us" button. Suggestions, feedbacks and messages will be automatically recorded and forwarded to the administrators.

As mentioned before, terms in the termbase are selected from both electronic documentation sources and hardcopy, including books, textbooks, pamphlets, journals, magazines, dictionaries, glossaries, standards and databases. Due to the fast-developing nature of CA, it is also necessary to assign Chinese equivalent terms to newly incorporated concepts according to term-formation methodology (J. C. Sager 1997). Notice should also be paid to whether the source for terms was originally written in Chinese or a translation from other languages. If it is a translation, the equivalents translated should be checked for authenticity based on the comparison with the identified terms in the document in the original language (Pavel & Nolet 2001). Potential terms are then submitted to careful examination and confirmation by both language experts and field experts.

As CA is a young field of study, it is highly possible that terminology in this field displays both "a certain degree of dynamism" and "inconsistency of usage (A. Kilyeni *et al*. 2012)". Therefore, for some terms, a "variants and a synonyms" data category should be included, in which the variants comprise geographical variants (e.g. variants used in Taiwan and the mainland China), syntactic variants, abbreviations and acronyms, and the latter, synonymous terms. Remarks concerning the variants and synonyms should be noted down. For instance, for the English expression "start the engine", the two Chinese equivalents should be mentioned: 启动发动机 and 开车, with the observation that the latter is preferred by instructors and encountered more often in teaching or tutorial materials like textbooks. At this stage, it is important to hold forums and conduct discussions involving language and field experts to sort out the terms, i.e. whether they are recommended, cautioned against, misused or avoided. Although the current version of termbase supports searches in simplified Chinese (variants in traditional Chinese are given but not searchable) and English, an updated version will enable searches in traditional Chinese.

In addition to the definition and variants provided, some headwords are explained with pictures for they "provide visual support for the verbal description of the semantic content of linguistic items (B. Svensén 2009, 298)" and therefore more vivid. One example is the term Wing. *Longman Dictionary of Contemporary English* (2014) defines it as "one of the large flat parts that stick out from the side of a plane and help to keep it in the air". However, a termbase user who reads this definition may wonder what a *wing* looks like, where it is in a plane, etc. As a result, when a user conducts a search for the term *wing*, the search result will be like Figure 3. With the help of a picture, the user might get a clue of the information that he or she is in need of.

Fig. 3 – *The search result for wing in the termbase*

Moreover, if a term has more than one meaning in this field (e.g. in case of polysemy), this term should be recorded in a separate entry for each concept it denotes.

5. Subdivision of the Field of CA

According to S. Pavel & D. Nolet (2001), division of a field into smaller units is a basic principle in terminology work. Accordingly, concepts in the current termbase of CA should be organized based on the logical structure of knowledge that offers its users with a clear and coherent picture of the field. Due to the large number of concepts to be analysed for the termbase, a decision has been arrived that the field of CA can be divided into five main subfields[4], they are 1) civil aircraft, 2) air transport geography, 3) flight oper-

4 Due to statistical needs, ICAO (2009) has classified the commercial air transport services into eight main subfields, they are 1)commercial air transport services, 2)general aviation, 3)airport services, 4)air navigation services, 5)civil aviation manufacturing, 6)aviation

ation, 4) quality and safety management and 5) ground service. Each main subfield can also be further divided, e.g. civil aircraft is a subfield consisting of two parts, i.e. civil aircraft design and manufacturing; air transport geography is made up of four parts including transportation and geography, transportation and spatial structure, transportation terminals as well as transport planning and policy; flight operation can be divided based on the checklist into eight parts, i.e. before starting, engine starting, before taxiing, lining up, in flight, approaching, landing and shut down; quality and safety management is divided into five parts, i.e. principles of quality management, information sharing, top management support, risk management and safety assurance as well as emergency planning and incident command; ground service consists of four parts which are the supply of oil and water, passenger boarding, baggage handling and waste disposal.

Given the cross-domain feature of CA, it would be suggested to indicate the primary sub-domain to which a concept belongs. In addition, as the termbase consists of five subfields and each subfield can be further divided, the subfields shown of the research query is depended on the choice made by the user. For instance, when a user searches for a term via the advanced search function and chooses the second subfield (air transport geography), the search result will only yield the terms found in this subfield, and when the user chooses the third (flight operation) and fourth (quality and safety management) subfields, a search result of these two subfields will be shown.

6. Division of Roles and Functions of the Termbase

As can be seen from the previous sections of the article, this termbase mainly consists of two roles, i.e. registered user(s) and administrator(s). After registration, a user can search for terms, symbols, markings and signs in the database, and can have personalized services, e.g. add term entries to a termlist which is accessible by himself or herself only. Besides, a registered user can point out inaccuracies or propose new terms, suggestions, and so on by sending messages.

training, 7)maintenance and overhaul and 8)regulatory functions and other activities. We adopt, however, the above-mentioned classification of five subfields, because it is more CA-oriented.

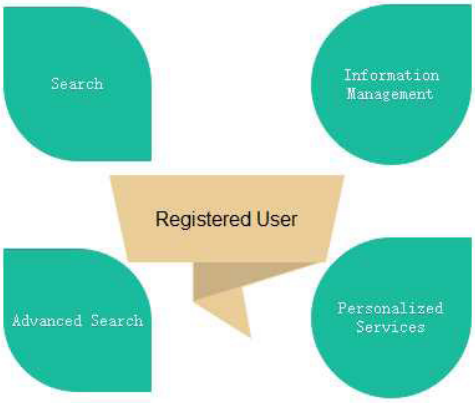


Fig. 4 – Functions of the termbase for a registered user

Similarly, in the background of the system, staff are needed to manage term entries, add new terms, modify the entries, manage the registered users, etc. Such a person is a system administrator. Figure 5 is an illustration of the role and responsibilities of a system administrator.

Fig. 5 – Role and responsibilities of a system administrator

7. Benefits of the Termbase

Undoubtedly, such a termbase in CA as presented above would offer valuable benefits to a variety of users. For a start, Chinese CA practitioners, transportation management departments and other airline groups will be able to benefit from its precise and unambiguous terminology for a successful professional communication. On the other hand, the termbase would provide assistance for academic staff involved in teaching CA and other related fields (e.g. aviation, engineering, communication, etc.) as well as students who need to familiarize themselves with domain terminology. Thirdly, the termbase will be a reliable tool for translators and interpreters. Moreover, it is beneficial for an effective exchange of knowledge and information between Chinese and foreign field specialists in their joint efforts for a sustainable implementation of information management and quality standards. Fifthly, a systematic organization of CA terminology according to logical and ontological relations between concepts allows users to obtain a coherent and clear "concept network" of the field. Last but not least, the termbase ensures that information can be easily disseminated and accessed by a great number of users at any time.

8. Conclusion

A termbase is particularly important for a field which requires precise communication like CA. In addition, a flexible CA termbase would provide an excellent solution to present-day global efforts towards the harmonization and the subsequent standardization of the subject-field terminology.

The concepts and methods proposed in this paper can be regarded as the foundation for the creation of a termbase in CA or other related fields. The creation of such a CA termbase starts from the analysis and selection of terms which should consist of not only words, but also symbols, markings and signs. A variety of functions are provided through the division of roles. It is hoped that the multi-functions and management of this termbase are the first step towards the development of a comprehensive and widely accepted standard of concepts and terms in the field of CA that will greatly improve the efficiency of field communication both at a national and international level.

References

AGARD/NATO. 1980. *Multilingual Aeronautical Dictionary*. Brussels : North Atlantic Treaty Organization-Advisory Group for Aerospace Research and Development.

Annex 14 to the Convention on International Civil Aviation. Volume I. Aerodrome Design and Operations. 2018. 8th Edition. Canada : ICAO (International Civil Aviation Organization).

Babić, Stjepan. 1990. "Postanak Novih Naziva (Origin of New Words)." *Naučni Sastanak Slavista u Vukove Dane* 18 (1): 31-37.

Cabré, M. Teresa. 1999. *Terminology : Theory, Methods, and Applications*. Amsterdam, Netherlands : John Benjamins.

Crane, Dale. 2012. *Dictionary of Aeronautical Terms*. Newcastle : Aviation Supplies & Academics, Inc.

Feng, Zhiwei. 1997. *An Introduction to Modern Terminology*. Beijing : Language and Culture Press.

Guo, Bozhi, & Chen, Yingchun (eds). 2011. *Technical Terms of Civil Aircraft*. Beijing : Aviation Industry Press.

Guo, Heping. "On Philosophical Terms : A Cause concerning the Development of Philosophy and Society". *Journal of Yunnan Normal University (Humanities and Social Sciences Edition)* 38(3): 101-104.

He, Daode & Zhou, Ruilian. 1997. *A New English-Chinese Civil Aviation Dictionary*. Beijing : China Civil Aviation Publishing House.

Hu, Tong, Huang, You, & Sun, Jiandong. 2016. *Concise English-Chinese Dictionary for Pilots*. Beijing : China Civil Aviation Publishing House.

Huang, Dexian. 2017. "Transparent Translation of Civil Aviation Terminology". *Journal of Beijing University of Aeronautics and Astronautics (Social Sciences Edition)* 30(1): 108-114.

Huang, Min. 2015. *An English-Chinese Dictionary of Aeronautic Technology Abbreviations and Acronyms*. Shanghai : Shanghai Jiao Tong University Press.

Infoterm. 2005. Guidelines for Terminology Policies. Formulating and Implementing Terminology Policy in Language Communities". Accessed Jan. 6, 2019, through http://www.unesco.org/new/en/communication-and-information/resources/publications-and-communication-materials/publications/full-list/guidelines-for-terminology-policies-formulating-and-implementing-terminology-policy-in-language-communities/.

ISO 7001 Graphical Symbols - Public Information Symbols. 2007. Geneva : Inter-national Organization for Standardization.

ISO 12620 Management of Terminology Resources- Data Category Specifications. 2019. Geneva : Inter-national Organization for Standardization.

Jones, David. 2003. *The Aviation Dictionary for Pilots and Aviation Maintenance Technicians*. Connecticut : Jeppesen Sanderson, Inc.

Kilyeni, Annamaria, Ciobanu, Georgeta, & Palea, Adina. 2012. "Designing a Database for Landscape Architecture Terminology". *Procedia- Social and Behavioral Sciences* 46 : 4666-4671.

Kumar, Bharat, De Remer, Dale, & Marshall, Douglas. 2002. *An Illustrated Dictionary of Aviation*. New York : McGraw-Hill Education.

Ledrew, Shirley. 1997. Terminology, Semantics and Lexicography. In *Terminology : A Practical Approach*, ed. by Dubuc Robert, 23-34. Portobello : Linguatech.

Liang, Bingwen. 2000. *The English-Chinese Dictionary of Aeronautical and Astronautical Engineering.* Beijing : Commercial Press.

Lipka, Leonhard. 1992. *An Outline of English Lexicology*. (2nd ed.). Tübingen : Max Niemeyer Verlag.

Liu, Deyi. 2005. *An Introduction to Civil Aviation*. Beijing : China Civil Aviation Publishing House.

Liu, Xin & Zhou, Jian. 2011. "A Brief Discussion on the Terminology of Air Traffic Control". *Science & Technology Information* 18(2): 174.

Milić, Mira. 2015. "Creating English-based Sports Terms in Serbian-Theoretical and Practical Aspects". *Terminology* (21)1 : 1-22.

Pavel, Silvia, & Nolet, Diane. 2001. *Handbook of Terminology*. Quebec : Translation Bureau, Minister of Public Works and Government Services Canada.

Longman Dictionary of Contemporary English. 2014. Pearson Education ESL. Beijing : Foreign Language Teaching and Research Press.

Pearson, Jennifer. 1998. *Terms in Context*. Amsterdam/Philadelphia : John Benjamins.

Pu, Jianjun & Chen, Aisha. eds. 2012. *Civil Aviation English*. Chengdu : Southwest Jiaotong University.

Review of the Classification and Definitions used for Civil Aviation Activities. 2009. Canada : ICAO (International Civil Aviation Organization).

Sager, Juan, C. 1997. "Term Formation". In Wright, Sue, Ellen, &Budin, Gerhard (eds.). *Handbook of Terminology Management. Vol. 1. Basic Aspects of Terminology Management*, 25-42. Amsterdam/Philadelphia : John Benjamins.

Shekhar, R.K.C. , & Shekhar, S.K.C. 2005. *Dictionary of Civil Aviation*. Delhi : Gyan Books.

Svensén, Bo. 2009. *A Handbook of Lexicography*. Cambridge : Cambridge University Press.

Terms in Aeronautical Science and Technology. 2004. China National Committee for Terms in Sciences and Technologies. Beijing : The Commercial Press.

Wen, Jun & Li, Peijia. "On the Translation Methods of Aerospace Terms". *Journal of Guangdong University of Foreign Studies* (22)3 : 27-32.

Yu, Tao. 2011. *English for Civil Aviation Service*. Beijing : China Civil Aviation Publishing House.

Zhang, Fan. 2017. "On the Failures of Civil Aviation Cabin Safety Language and Their Counteraction". *Journal of Civil Aviation Flight University of China* 28(3):34-38.

Zhang, Huaixing. 2011. *An English-Chinese Dictionary of Civil Aviation*. Beijing : China Civil Aviation Publishing House.

Zhou Qihuan. 2002. "A General Analysis about the Difference of Aeronautical Terms Used in Both Sides of the Taiwan Straits". *Terminology Standardization & Information Technology*. 7(2): 8-11.

Zhou Qihuan. 2010. "Investigation on Normalization of Chinese Civil Aviation Scientific Terms". *Journal of Civil Aviation University of China* 28(4):41-46.

Zhu, Shixing. 2013. *An English-Chinese Dictionary of Civil Aviation Abbreviations and Acronyms*. Beijing : China Civil Aviation Publishing House.

Résumé

Actuellement, une communication précise dans un domaine particulier est impossible sans une terminologie appropriée. Il en va de même pour l'aviation civile, qui est un domaine interdisciplinaire qui se développe rapidement et qui a besoin d'une extrême précision. Basé sur une analyse linguistique de termes dans le domaine de l'aviation civile (CA), cet article explore les aspects clés liés à la construction et à la compilation d'une base terminologique bilinulaire (base de données terminologiques) dans ce domaine et vise à souligner le besoin impératif de un tel outil pour favoriser l'échange efficace de connaissances et d'informations dans un cadre professionnel. Les principes de conception de cette base terminologique couvrent trois aspects essentiels : la sélection de termes et d'images, les options d'affichage ainsi que

la répartition des rôles et des fonctions. La pratique et les concepts peuvent également être étendus pour créer des bases terminologiques pour d'autres langues et domaines.

Validating a SKOS representation of a manually developed terminological resource. A case study on the quality of concept relations

Christian Lang*, Karolina Suchowolec**,
Matthias Wischnath***

*Leibniz-Institut für Deutsche Sprache
Augustaanlage 32
D-68165 Mannheim
lang@ids-mannheim.de
**TH Köln
karolina.suchowolec@th-koeln.de
***TH Köln
matthias_gabriel.wischnath@smail.th-koeln.de

Abstract. In our paper, we present a case study on the quality of concept relations in the manually developed terminological resource of *grammis*, an information system on German grammar. We assess a SKOS representation of the resource using the tool qSKOS, create a typology of the issues identified by the tool, and conduct a qualitative analysis of selected cases. We identify and discuss aspects that can motivate quality issues and uncover that ill-formed relations are frequently indicative of deeper issues in the data model. Finally, we outline how these findings can inform improvements in our resource's data model, discussing implications for the machine readability of terminological data.

1. Introduction

Recently, there has been a trend towards increasing machine readability of terminological data. This trend amounts to a paradigm shift in the field, as the target audience of terminological resources has traditionally been human users such as translators or technical writers, whereas machine readability has been the focus of ontology building (cf. Drewer *et al*. 2017:22). These

developments also imply that the formal aspects of concept relations in traditional terminological resources require special attention, because relations need to satisfy stricter constraints that are posed on properties in formal and semi-formal representations. This is particularly challenging in manually developed terminological resources for human users.

In our project, we address the trend outlined above and deal with an automatically retrieved SKOS representation[1] of the manually developed, human-oriented terminological resource of *grammis*.[2] Following Mader *et al*. (2012), we do an automated quality assessment of this SKOS representation using the tool qSKOS[3] and focusing on the quality of concept relations. In this paper, we present the results of this quality assessment and share insights garnered from a qualitative analysis of these results. We begin by giving background of this case study. We then describe our methodology and present our preliminary results. Finally, we discuss selected cases from a qualitative analysis with regard to their implication for our resource.

2. Previous Work

Grammis is an information system on German grammar hosted at the Leibniz-Institute for the German Language (IDS) (cf. Schneider/Schwinn 2014). As discussed in Suchowolec *et al*. (2017), we are currently revising *grammis*' terminological resource according to terminological standards and best practices. One goal of the revision is to provide human users with an overview of the diversity of grammatical theories by integrating concepts from different theoretical backgrounds. Furthermore, we seek to add a machine-readable layer to the resource and publish it as Linked (Open) Data. As a proof of concept, Suchowolec *et al*. (2017, 2019) showed that *grammis*' terminological data can be represented as a SKOS vocabulary using the D2RQ platform.[4] However, Suchowolec *et al*. (2019) pointed out that *grammis*' terminological resource had not been developed with a future (semi-) formal representation in mind and, hence, the correctness of the SKOS representation needs to be validated. This pertains in particular to the well-formedness of concept relations.

1 Strictly speaking, a SKOS-XL representation. For reference, see https://www.w3.org/TR/skos-reference/skos-xl.html.
2 https://grammis.ids-mannheim.de/.
3 https://github.com/cmader/qSKOS.
4 http://d2rq.org/.

Our case study addresses this issue and deals with a formal validation of *grammis*' SKOS representation in order to explore the quality of conceptual relations in a manually developed, human-oriented, onomasiological, and descriptive terminology resource.

3. Related Work

SKOS has traditionally been used to represent controlled vocabularies such as thesauri. Van Assem *et al*. (2006) proposed a general three-step approach for converting existing thesauri to a SKOS representation. They applied this approach in several case studies to vocabularies such as IPSV, GTAA, and MeSH. Others built on this approach later on, for example Neubert (2009) and Zapilko *et al*. (2012) in their conversion of STW and TheSoz respectively. More recently, Chiarcos et at. (2016) converted BLL Thesaurus to a SKOS representation as an intermediate step towards a Linguistic Linked Open Data resource.

Converting *grammis*' terminological data into a SKOS vocabulary, Suchowolec *et al*. (2019) follow the general approach as described in van Assem *et al*. (2006). However, some issues reported in the studies mentioned above, such as identifying concepts in a semasiological thesaurus or mapping compound concepts (van Assem *et al*. 2006 : 100 ; section 4. "Modelling Issues" in Zapilko *et al*. 2012), are thesaurus-specific. Furthermore, the SKOS-XL extension of the original SKOS specification has already addressed other reported issues that also pertain to a traditional onomasiological terminology resource, such as the need for a mapping mechanism between preferred and non-preferred terms.

With respect to quality assessment of SKOS vocabularies, Mader *et al*. (2012) identify quality issues in 15 SKOS vocabularies (AGROVOC, DBpedia, and GTAA i.a.) through automatic formal validation using Mader's tool qSKOS. Suominen/Hyvönen (2012) analyze 14 SKOS vocabularies using criteria from, amongst others, qSKOS and find that many of them contain structural errors.

4. Method

In this section, we describe the database and our methodology. Our approach consists of four steps : first, we convert our terminological database into RDF triples. Second, we run a qSKOS quality assessment. Third, we

annotate the results of the qSKOS quality report; and, fourth, we perform a qualitative analysis of the quality issues regarding concept relations.

4.1. Database and Conversion

Fig. 1 – *Graph-object depicting concepts and concept relations of grammis' terminological database*

Our original database is the terminological resource of *grammis* (cf. Schneider 2007). According to Zeng's criteria to categorize knowledge organization systems (2008:161),[5] *grammis'* terminological resource can be considered a thesaurus.

In our case study, we use *grammis'* test environment as described in Suchowolec *et al*. (2017, 2019). First, we update its content to the latest version.

5 "Eliminating ambiguity, controlling synonyms, establishing relationships: hierarchical, establishing relationships: associative" (Zeng 2008:161).

The current resource contains 4,534 terms, 1,425 concepts, 1,056 instances of generic hierarchical relations (BT), 586 instances of partitive hierarchical relations (BTP), and 2,555 instances of associative relations (RT). Fig. 1 shows a visualization[6] of the test environment and gives a first indication of possible quality issues (see sec. 4.2).

In a second step, we create an RDF dump from the test environment using the D2RQ dump-rdf tool. The resulting RDF dump consists of 41,166 triples.

4.2. qSKOS Quality Assessment and Preliminary Results

We perform an automatic validation of the RDF dump created in *grammis'* test environment using version 2.0.3 of qSKOS[7] and following Mader *et al.*'s approach (2012). qSKOS "goes beyond SKOS model integrity. It implements checks for additional quality criteria that were derived from existing literature and guidelines […]" (Mader/Haslhofer 2013 :9). This feature is of particular interest for *grammis*, as its revision is ongoing. Hence, the validation results can serve as "quantitative quality indicators" (Mader *et al.*: 2012 :232). In particular, they can indicate quality problems in the instantiation of concept relations in the resource. As Mader/Wartena (2014) show, an automatic validation is often the method of choice to spot quality issues due to the size of the RDF resource, i.e. the number of triples.

Table 1 shows an excerpt of the summary of the qSKOS validation. The full report also indicates some issues regarding labels (such as Missing Labels). However, since we focus on conceptual relations in this case study, we do not discuss these label issues here. We also disregard the issues concerning Missing Out-Links and Missing In-Links because we are working in a closed local test environment.

Summary of Quality Issue Occurrences:	
Orphan Concepts	FAIL (7)
Disconnected Concept Clusters	FAIL (2)
Valueless Associative Relations	FAIL (313)

6 We created the visualization with the open-source software gephi 0.92 (Bastian *et al.* 2009) using the ForceAtlas algorithm. Nodes represent concepts; the edges between concepts represent concept relations.

7 qSKOS has been further developed and integrated into PoolParty SKOS Quality Checker (https://qskos.poolparty.biz/login). While PoolParty SKOS Quality Checker allows for formal validation of data in the cloud, qSKOS can be installed as a local client, which is preferred in our setting.

Hierarchical Redundancy	FAIL (91)
Relation Clashes	FAIL (43)

Tab. 1 – *Excerpt of quality issues summary (focus on concept relations)*

The following descriptions of the quality issues shown in table 1 follow the wiki on Mader's GitHub-page (Mader 2014).

Orphan Concepts (OC) are isolated concepts that are not connected to other concepts via associative or hierarchical relations. The seven orphan concepts in our database can be seen at the bottom of the visualization in Fig. 1.

Disconnected Concept Clusters (DCC) are concept clusters that are not connected to other concept clusters. Fig. 1 shows two clusters ; the main cluster accounting for the majority of concepts and one small cluster in the upper left corner.

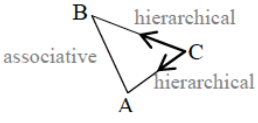


Fig. 2a – *Valueless Associative Relations*

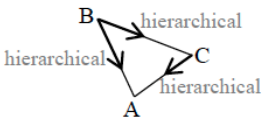


Fig. 2b – *Hierarchical Redundancy*

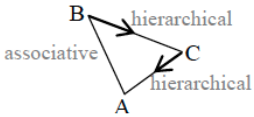


Fig. 2c – *Relation Clashes*

Valueless Associative Relations (VAR) are associative relations (RT) between two concepts (A, B) that share the same mother concept (C), i.e. are siblings ; Fig. 2a. shows an example of VAR.[8]

Hierarchical Redundancy (HR) refers to a constellation of relations where a pair of concepts (A, B) is directly hierarchically related and there also exists a hierarchical path through a concept C that connects A and B ; see Fig. 2b for an example of HR. Note that according to example 34 in the SKOS reference (Miles, Bechhofer 2009 :s.p.), *skos :broader* and *skos :narrower* are intransitive properties. However, they are sub-properties of *skos :broaderTransitive* and *skos :narrowerTransitive* respectively.[9] As Mader (2014 :s.p.) puts it : "This, in fact, leaves it up to the user to interpret wheter [*sic*] a vocab-

8 According to ISO 25964-1 :2011(E) (p. 63), it is not necessary to connect co-hyponyms with an associative relation.

9 This means that according to the SKOS reference "<A> skos :broader <B> . <B> skos :broader <C> ." entails "<A> skos :broaderTransitive <B> . <B> skos :broaderTransitive <C> . <A> skos :broaderTransitive <C> ." (example 35, Miles, Bechhofer 2009 :s.p.).

ulary's hierarchical structure is seen as transitive or not. In the former case, this check can be useful."

Among the issues shown in table 1, **Relation Clashes (RC)** is the only issue category that is associated with a SKOS integrity condition – namely S27 : "skos :related is disjoint with the property skos :broaderTransitive" (Miles, Bechhofer 2009 :s.p.) – and concerns a clash between associative and hierarchical relations.[10] An example of RC is illustrated in Fig. 2c.

In the remainder of this paper, our main focus will be on Valueless Associative Relations, Hierarchical Redundancy and Relation Clashes.[11]

4.3. Further Processing of the qSKOS Results

In order to gain deeper insight into the results, further processing of the report data is necessary. The procedure consists of three steps.

Valueless Associative Relations, Hierarchical Redundancy and Relation Clashes all involve three concepts. The report, however, only shows two concepts (cf. Fig. 3). Thus, we first have to identify the respective third concept. We query the database and facilitate this task by also applying a visualization tool that was developed within the project (cf. Suchowolec *et al*. 2019 for a detailed description of the tool).

```
(http://localhost:2020/TB_KONZEPT/133, http://localhost:2020/TB_KONZEPT/135)
(http://localhost:2020/TB_KONZEPT/1383, http://localhost:2020/TB_KONZEPT/163)
(http://localhost:2020/TB_KONZEPT/252, http://localhost:2020/TB_KONZEPT/135)
```

Fig. 3 – *excerpt of qSKOS quality report ; each row represents one issue, note that the report only gives two concepts per issue*

As a result of this step, we can represent every instance of the three quality issues by a concept triangle.

10 "<A> skos :broader <B> ; skos :related <C> . <B> skos :broader <C> ." (example 27, Miles, Bechhofer 2009 :s.p.); the following inference can be drawn : "<A> skos :broaderTransitive <C>; skos :related <C> ." (example 28, Miles, Bechhofer 2009 :s.p.).

11 As for Orphan Concepts (OC) and Disconnected Concept Clusters (DCC), these two issues become apparent from Fig. 1. With only seven instances, OC (at the bottom of Fig. 1) are relatively rare in our data and, i.a., involve test concepts. Regarding DCC, Fig. 1 shows two clusters ; one consists of the main part of our data, the second one consists of mere three concepts that are not part of the grammatical core terminology (see the upper left part of the visualization). The issues of both categories have been fixed.

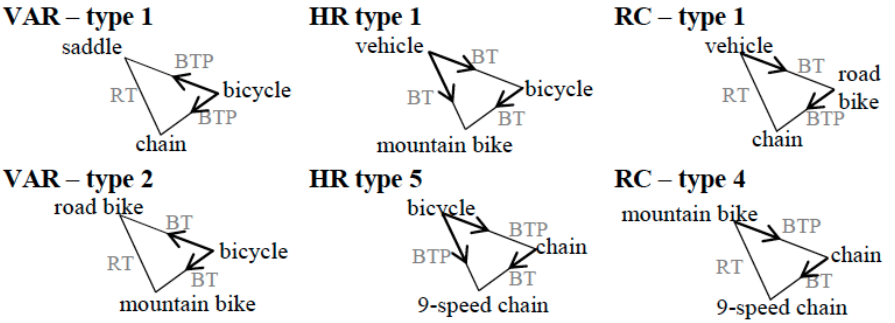


Fig. 4 – *Most frequent subtypes of each quality issue with theory-driven example concepts; RT=associative relation, BT=generic hierarchical relation, BTP=partitive hierarchical relation*

In a second step we annotate the hierarchical relations between the concepts of the triangles. This is necessary, because the SKOS representation – unlike our original database – does not differentiate between generic (BT) and partitive hierarchical relations (BTP). The result of this annotation is a preliminary typology with up to seven subtypes per quality issue.

As a third step, we add theory-driven example concepts to each of the subtypes. This helps us to refine the typology, to spot candidates for ill-formed relations, and to create references for the subsequent qualitative analysis of the issues. Figure 4 shows the two most frequent types for each quality issue with theory-driven example concepts added. Note that the example concepts come from the general domain of *bicycle*, which we assume to be 1) easy to follow for a general audience and 2) prototypical for relation types BTP and BT, giving us a top-down priming.

4.4. Qualitative Analysis

In accord with Mader *et al.* (2012:232) who point out that structural quality issues require further inspection, we perform a qualitative analysis of the results of the qSKOS report after they were further processed. We analyze each quality issue on a case-by-case basis and look for motivating factors behind the ill-formed relations. We believe that an in-depth analysis of the quality issues can provide insight into the challenges of converting a manually developed, human-oriented, onomasiological, and descriptive terminology resource into a (semi-) formal representation.

5. Results

In this section, we first present an overview of the quantitative results and then discuss selected cases from the qualitative analysis.

5.1. Quantitative Results – Overview

On the basis of the typology described in sec. 4.3, we are able to give a more detailed presentation of the quality issues in our database. Table 2 shows the distribution of the subtypes of different quality issues. In each category, we encounter instances where the third concept cannot be determined (cf. column "C unclear").

Issue/type	Σ	1	2	3	4	5	6	C unclear
VAR	**313**	100	139	9	38	10	4	13
HR	**91**	23	9	9	17	6		27
RC	**43**	8	1	13	3			18
Σ	**447**							**58**

Tab. 2 – *Distribution of different types of each quality issue*

VAR, the most common issue category, accounts for 70% of the total issues (HR: 20%, RC: 10%). Due to the high share of VAR, 80% of the 447 issues reported in total involve associative relations (VAR and RC combined).

We find that cases are unevenly distributed across all issue categories. In VAR, type 1 and type 2 account for 76% of the instances (cf. Fig. 4 for more information). In HR, the most frequent subtypes, type 1 and type 4, make up approx. 44% of the instances. Finally, in RC, type 1 and type 3 account for approx. 48% of the cases.

Interestingly, 21 out of 43 RC show a combination of generic and partitive relations, which makes the ratio of mixed relations in RC considerably higher than in the other issue categories.

5.2. Insights from the Qualitative Analysis – Selected Cases

In the following sections, we present selected cases from each quality issue from our qualitative analysis. Generally, we find two types of quality issues: 1) obvious mistakes, i.e. *un*motivated ill-formed relations, and 2) *motivated* ill-formed relations, in the sense that they code valuable information

and/or point to deeper underlying issues. The following cases illustrate some of the (linguistic) motivations behind ill-formed relations.[12]

6. Valueless Associative Relations

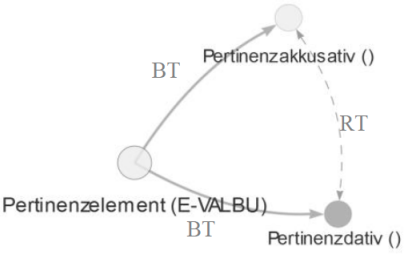


FIG. 5a – *VAR type 2: Pertinenzelement ('pertinence element')*

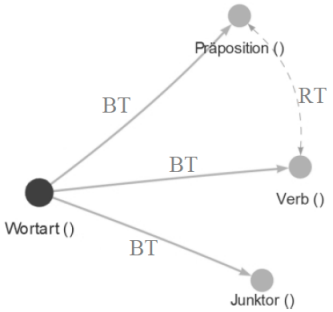


FIG. 5b – *VAR type 2: Wortart ('Part of Speech')*

As outlined in sec. 4.2, VAR refers to co-hyponyms connected via an associative relation. *Prima facie*, the redundancy of the associative relation between two siblings seems to be evident as their connection follows from sharing the same mother. We find instances in our data that confirm this initial assumption. We see an instance of this in Fig. 5a: As *Pertinenzelement* (*'pertinence element'*) has only two hyponyms, the existing link via an associative relation is superfluous.

However, we also find cases where an associative relation between siblings is motivated. For example, in Fig. 5b, qSKOS flagged the associative relation between *Verb* (*'verb'*) and *Präposition* (*'preposition'*) as VAR because both share the same mother *Wortart* (*'part of speech'*, *'POS'*). We believe, however, that in this case the flagged RT codes additional information. When taking a syntactic perspective, this becomes apparent, as POS function as building blocks of phrases and sentences. From this point of view, there is a special connection between verbs and prepositions, as some verbs require (govern) a prepositional object (e.g., *denken **an***, *'to think **of**'*). The flagged RT represents this special connection and reflects the fact that *verb* and *preposition* are more closely linked than, for example, *verb* and *'Junktor* (*'conjunction'*).

12 We limit ourselves to selected motivating factors as an exhaustive presentation would go beyond the scope of this paper.

6.2.1. Hierarchical Redundancy

As we indicated in sec. 4.2, HR is related to the transitivity status of the hierarchical relations in a data model.

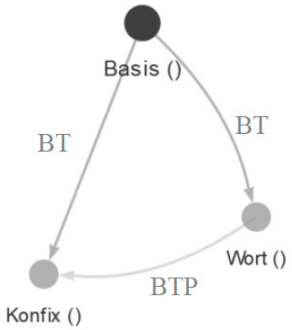


Fig. 6 – *HR type 3 : Basis (‘base’)*

Fig. 6 shows a case of HR that involves concepts pertaining to the domain of word formation. At first glance, the flagged BTP seems superfluous as it connects two concepts that share the same mother. However, upon further analysis, the ill-formed BTP points to a deeper underlying issue.

Basis (‘*base*’) refers to different elements that function as a basis for derivation. Both *Wort* (‘*word*’, e.g. ***freund****-lich*, *friend-ly*) and *Konfix* (‘*confix*’, *combining form, e.g.* ***polit****-isch, polit-ical*) can function as building blocks of derivation. Hence, the two concepts are modelled as co-hyponyms of *base.* Additionally, *confix* is modelled as the meronym of *word*. The result is a constellation of relations where *confix* is both directly and indirectly connected to *base,* i.e., a case of HR. If we assume that the hierarchical relations in our database are transitive (cf sec. 4.2), the partitive hierarchical relation connecting *word* and *confix* seems redundant.

However, the flagged relation in this case is associated with the fact that words, obviously, are not only building blocks in derivation, but also the result of this process. This means that a word containing a confix can again serve as base for another derivation process (e.g., [[***Zimper****-lich*]$_1$*-keit*]$_2$, ‘*squeamishness*’). The ill-formed BTP reflects this aspect.

6.2.2. Relation Clashes

As we indicated in sec. 5.1., a considerable amount of RC cases involves a combination of partitive and generic hierarchic relations.

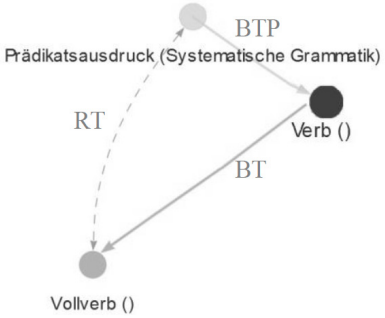


Fig. 7 – *RC type 1 : Prädikatsausdruck ('predicate expression')*

In Fig. 7, a partitive hierarchical relation connects *Prädikatsausdruck* ('*predicate expression'*) to *Verb* ('*verb'*) and there is a generic hierarchical relation between *verb* and *Vollverb* ('*full verb*'). Also, an associative relation connects *full verb* and *predicate expression*. This constellation represents a clash between associative and hierarchical relations and violates SKOS integrity condition S27 (cf. sec. 4.2).

The SKOS representation does not replicate the distinction between generic and partitive hierarchical relations that our original database makes. However, Winston *et al*. (1987:434-435) show that mixed relations, in fact, allow for valid syllogisms, albeit under certain conditions. Therefore, if we assume the relations in our database are transitive, we can draw an analogous inference to the one presented in sec. 4.2 (the same applies to the example shown in Fig. 6). In that case, the associative relation is, in fact, superfluous as its function to stress the connection between two concepts pertaining to different linguistic schools (*predicate expression* originating from Zifonun *et al*. (1997) and *full verb* being a widespread concept not associated with any particular theory) is rendered redundant by the inference.

6.1. Discussion

In essence, after a qualitative analysis we find that the results of our qSKOS assessment are ambiguous. Some of the quality issues constitute unmotivated mistakes we can remedy by simply removing or replacing the flagged relation. Others, however, are motivated and either code valuable information and/or point to deeper underlying issues. These cases of motivated issues indicate aspects that we have to consider in order to improve our database.

The discussions of HR and RC (sec. 6.2, 6.3) indicate that one aspect we have to address is the transitivity status of the relations in our database.[13] Moreover, the high share of RTs involved in the quality issues (sec. 6.1, 6.3) suggests that the limited number of relation types in our data is possibly insufficient to model the contents of our field adequately (cf. Lehmann 1996). We find that in our database RTs have a generic function and cover a wide variety of cases where the other two relation types do not apply. This comes as no surprise; associative relations are sometimes deemed less central than hierarchical relations in terminology (Drewer *et al*. 2017:18) and Dextre Clark (2016:142) claims that "[u]ntil about 20 years ago, the only purpose of the associative links (RT) in a thesaurus was to help the indexer or searcher navigate the thesaurus and think of more terms to use […]." We believe that the high share of flagged RTs and their generic function in our database suggest the necessity to consider an extension of our inventory of relations, e.g. by means of a relation coding the syntactical function *government* in the case shown in Fig. 5b (see Hjørland 2016:151 on the introduction of domain-specific relations).

7. Summary

In our paper we presented the results of an automated quality assessment of a SKOS representation. The database of the representation was *grammis*, a manually developed terminological resource. Our focus was on issues regarding conceptual relations and we discussed selected cases from a qualitative analysis. We found qualitative issues in our database – unmotivated and motivated – and it is the latter that make us reconsider some aspects of our resource's structure. Above all, we need to address formal aspects of

13 Suominen/Hyvönen (2012:395) find that issues that violate SKOS integrity condition S27 (like RC does) are a very common occurrence in the vocabularies they analyzed and speculate that integrity condition S27 might be too strict.

the relations, restrictions on transitivity (in context of mixed partitive and generic hierarchies), and restrictions on the domain and range of relations. Furthermore, we need to consider an extension of our inventory of relations. The implementation of these measures would mean becoming more formal and, as a consequence, taking a step towards ontology building; thereby following what Mazzocchi *et al*. (2007:199) call the "trend towards the refinement of relational semantics" in thesauri.[14] Taking this step is a very labor-intensive process[15] but could improve the degree of internal structural consistency of a thesaurus (Mazzocchi *et al*. 2007:200).

In sum, a part of the Terminology community argues for the use of existing terminology resources in the context of ontology and AI applications, and calls for an increase in machine readability of terminological data. Our case study suggests that the similarity of resources based on concept systems and ontologies may be rather conceptual; a simple application of more strict formalisms to an already existing human-oriented resource might not lead to satisfactory results. Further, as mentioned above, some important aspects need to be addressed *a priori* from an early stage of the resource draft. All in all, the problems we faced in our case study are similar to those faced previously in the Thesaurus community, which can be an important source of information for current trends in the Terminology community.

References

van Assem, Marc, Veronique Malaisé, Alistair Miles, and Guus Schreiber. 2006. "A Method to Convert Thesauri to SKOS." In *The Semantic Web: Research and Applications. 3rd European Semantic Web Conference, ESWC 2006 Budva, Montenegro, June 11-14, 2006 Proceedings*, edited by York Sure and John Domingue, 95-109. Berlin/Heidelberg: Springer.

Bastian, Mathieu, Sebastian Heymann, and Mathieu Jacomy. 2009. "Gephi: an open source software for exploring and manipulating networks." In *Proceedings of the Third International Conference on Weblogs and Social Media*, edited by Eytan Adar, Matthew Hurst, Tim Finin, Natalie Glance,

14 See the ongoing debate about the future prospects of thesauri in information retrieval (cf. the exchange between Dextre Clarke (2016) and Hjørland (2016).

15 Cf. Soergel *et al*. (2004) who describe the reengineering of the AGROVOC thesaurus into an ontology, see also Kless *et al*. (2015) for an analysis of the fundamental differences between thesauri and ontologies.

Nicolas Nicolov, and Belle Tseng, 361-362. Menlo Park : The AAAI Press. http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.

Chiarcos Christian, Christian Fäth, Heike Renner-Westermann, Frank Abromeit, Vanya Dimitrova. 2016. "Lin|gu|is|tik : Building the Linguist's Pathway to Bibliographies, Libraries, Language Resources and Linked Open Data." In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). 2016-05-23/2018-05-28, Portorož, Slovenia,* edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani *et al*., 4463-4471. Paris : ELRA. http://www.lrec-conf.org/proceedings/lrec2016/pdf/814_Paper.pdf.

Dextre Clarke, Stella G. 2016. "Origins and Trajectory of the Long Thesaurus Debate." *Knowledge Organization* 43 (3): 138-144.

Drewer, Petra ; François Massion, Donatella Pulitano. 2017. *Was haben Wissensmodellierung, Wissensstruktur, künstliche Intelligenz und Terminologie miteinander zu tun ?* Köln : Deutscher Terminologie-Tag e.V. http://dttev.org/DIT/168-was-haben-wissensmodellierung-wissensstrukturierung-kuenstliche-intelligenz-und-terminologie-miteinander-zu-tun.html.

Hjørland, Birger. 2016. "Does the Traditional Thesaurus Have a Place in Modern Information Retrieval ?" *Knowledge Organization* 43 (3): 145-159.

ISO 25964-1 :2011 (E). *Thesauri and interoperability with other vocabularies. Part 1 : Thesauri for information retrieval. First edition 2011-08-15.* Geneva : ISO.

Kless, Daniel, Simon Milton, Edmund Kazmierczak, and Jutta Lindenthal. 2015. "Thesaurus and Ontology Structure : Formal and Pragmatic Differences and Similarities." *Journal of the Association for Information Science and Technology* 66 (7): 1348-1366.

Lehmann, Christian. 1996. "Linguistische Terminologie als relationales Netz." In *Nomination – fachlich und gemeinsprachlich*, edited by Clemens Knobloch, and Burkhard Schaeder, 215-267. Opladen : Westdeutscher Verlag.

Mader, Christian. 2014. "Quality Issues". Last modified 15 Dec 2014. https ://github.com/cmader/qSKOS/wiki/Quality-Issues. Accessed October 31, 2019.

Mader, Christian, Bernhard Haslhofer, and Antoine Isaac. 2012. "Finding Quality Issues in SKOS Vocabularies." In *Therory and Practice of Digital Libraries. Proceedings of the Second International Conference, TPDL 2012, Cyprus,* edited by Panayiotis Zaphiris, George Buchanan,

Edie Rasmussen, and Fernando Loizides, 222-233. Berlin/Heidelberg : Springer. Preprint : http://arxiv.org/pdf/1206.1339v1.

Mader, Christian, and Bernhard Haslhofer. 2013. "Perception and Relevance of Quality Issues in Web Vocabularies." In *I-SEMANTICS 2013, Proceedings of the 9th International Conference on Semantic Systems*, edited by Marta Sabou, Eva Blomqvist, Tommaso Di Noia, Harald Sack, and Tassilo Pellegrini, 9-16. New York : Association for Computing Machinery. Preprint : http://eprints.cs.univie.ac.at/3720/.

Mader, Christian, and Christian Wartena. 2014. "Supporting Web Vocabulary Development by Automated Quality Assessment : Results of a Case Study in a Teaching Context." In *Workshop on Human-Semantic Web Interaction (HSWI'14), s.p.* http://hswi.referata.com/w/images/Hswi2014_paper_4.pdf.

Mazzocchi, Fulvio, Melissa Tiberi, Barbara Santis, and Paolo Plini. 2007. "Relational Semantics in Thesauri : Some Remarks at Theoretical and Practical levels." *Knowledge Organization* 34 (4): 197-214.

Miles, Alistair, and Sean Bechhofer, eds. 2009. "SKOS Simple Knowledge Organization System Reference." http://www.w3.org/TR/2009/REC-skos-reference-20090818/. Accessed Sep. 4, 2019.

Neubert, Joachim. 2009. "Bringing the 'Thesaurus for Economics' on to the Web of Linked Data." In *Proceedings of the WWW2 009 Workshop on Linked Data on the Web. Madrid, Spain 2009-04-20*, edited by Christian Bizer, Tom Heath, Tim Berners-Lee, and Kingsley Idehen, s.p. http://ceur-ws.org/Vol-538/.

Schneider, Roman. 2007. "A Database-driven Ontology for German Grammar." In *Data Structures for Linguistic Resources and Applications. Proceedings of the Biennial GLDV Conference 2007*, edited by Georg Rehm, Andreas Witt, and Lothar Lemnitzer, 305-314. Tübingen : Narr.

Schneider, Roman, Horst Schwinn. 2014. „Hypertext, Wissensnetz und Datenbank. Die Web-Informationssysteme Grammis und ProGr@mm." In *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache. Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*, edited by Franz Josef Berens and Melanie Steinle, 337-346. Mannheim : IDS Eigenverlag.

Soergel, Dagobert, Boris Lauser, Anita Liang, Frehiwot Fisseha, Johannes Keizer, and Stephen Katz. 2004. "Reengineering Thesauri for New Applications : the AGROVOC Example." *Journal of Digital Information* 4 (4): s.p., electronic only. https ://journals.tdl.org/jodi/index.php/jodi/article/view/112/111.

Suchowolec, Karolina ; Christian Lang, Roman Schneider, and Horst Schwinn. 2017. “Shifting Complexity from Text to Data Model.” In *Language, Data, and Knowledge. Proceedings of the First International Conference, LDK 2017, 2017-06-19/2017-06-20, Galway, Ireland*, edited by Jorge Gracia, Francis Bond, John P. McCrae, Paul Buitelaar, and Christian Chiarcos, 203-212. Cham : Springer.

Suchowolec, Karolina/Lang, Christian/Schneider, Roman. 2019) “An empirically validated, onomasiologically structured, and linguistically motivated online terminology – re-designing scientific resources on German grammar.” In : *International journal on digital libraries* 20(3): 253-268.

Suominen, Osma, Eero Hyvönen. 2012. “Improving the Quality of SKOS Vocabularies with Skosify.” In *Knowledge Engineering and Knowledge Management*, edited by Annette Teije, Johanna Völker, Siegfried Handschuh, Heiner Stuckenschmidt, Mathieu d’Acquin, Andriy Nikolov, Nathalie Aussenac-Gilles, and Nathalie Hernandez, 383-397. Berlin/ Heidelberg : Springer.

Winston, Morton E., Roger Chaffin, and Douglas Herrmann. 1987. “A Taxonomy of Part-Whole Relations.” *Cognitive Science* 11 : 417-444.

Zapilko, Benjamin, Johann Schaible, Philipp Mayr, and Brigitte Mathiak. 2012. “TheSoz : A SKOS Representation of the Thesaurus for the Social Sciences.” *Semantic Web – Interoperability, Usability, Applicability.* Online, s.p : http://www.semantic-web-journal.net/sites/default/files/swj279_2.pdf.

Zeng, Marcia Lei. 2008. “Knowledge Organization Systems (KOS).“ *Knowledge Organization* 35 (2/3) : 160-182.

Zifonun, Gisela, Ludger Hoffmann, and Bruno Strecker. 1997. *Grammatik Der Deutschen Sprache : Bd. 1-3*. Berlin : de Gruyter

Résumé

Dans notre article, nous présentons une étude de cas sur la qualité des relations conceptuelles dans la ressource terminologique de *grammis*, un système d’information sur la grammaire allemande, développée manuellement. Nous évaluons une représentation SKOS de la ressource à l’aide de l’outil qSKOS, nous créons une typologie des problèmes identifiés par l’outil et effectuons une analyse qualitative des cas sélectionnés. Nous identifions et discutons les aspects qui peuvent motiver les problèmes de qualité et découvrons que des relations mal formées sont souvent révélatrices de problèmes plus profonds dans le modèle de données. Enfin, nous expliquons comment ces résultats

peuvent contribuer à l'amélioration du modèle de données de notre ressource, en discutant des répercussions sur la lisibilité automatique des données terminologiques.

La technicité des termes : le *v-tech* comme paramètre d'évaluation

Federica Vezzani

Département d'études linguistiques et littéraires (DiSLL)
Piazzetta Gianfranco Folena, 1, 35137, Padoue - Italie
Université de Padoue
federica.vezzani@phd.unipd.it
http://www.dei.unipd.it/~vezzanif/

Résumé. Dans cette étude, nous proposons une perspective neuve du concept de poids des termes techniques en nous concentrant sur la notion de « technicité » comme propriété sémantique de l'unité linguistique elle-même. L'idée de base est que la valeur de technicité d'un terme est inversement proportionnelle à sa nature polysémique. Nous formalisons la formule *v-tech* et effectuons une évaluation expérimentale afin de 1) comparer la valeur *v-tech* avec d'autres mesures de *termhood* (termicité ou termitude) généralement calculées sur la fréquence d'occurrence des termes dans les collections, et 2) intégrer la formule *v-tech* dans le score d'un modèle de récupération de documents pertinents pour un travail de revue systématique dans le domaine médical.

1. Introduction

Cette étude s'inscrit dans le contexte de la terminologie computationnelle comme domaine d'étude récent qui vise à adopter des méthodes computationnelles et quantitatives afin de mener des recherches terminologiques et qualitatives (Bourigault *et al*. 2001, Foo 2012, Drouin *et al*. 2018). Dans la littérature, de nombreuses études sont principalement axées sur l'extraction automatique de termes à partir d'un corpus de documents spécialisés au moyen d'approches 1) linguistiques, 2) statistiques et 2) hybrides (Nakagawa 2001, Vu *et al*. 2008, Amjadian *et al*. 2018, Simon et Kešelj 2018, Sandoval *et al*. 2018). L'acquisition de termes liés et pertinents à un domaine spécifique de l'activité humaine est effectuée automatiquement à l'aide d'approches quantitatives importées du domaine de la recherche d'information : *Term*

frequency-Inverse Document Frequency (TF-IDF) (Salton et Yang 1973), *Mutual Information* (Church et Hanks 1990), T-Score (Church *et al*. 1991), C/NC (Frantzi *et al*. 1998).[1] En outre, des ressources spécifiquement conçues pour cette tâche ont été élaborées afin d'augmenter les performances d'extraction : TermoStat (Drouin 2003), BiTermEx (Planas 2012), TermEvaluator (Inkpen *et al*. 2016) et le récent projet MultiMedica (Sandoval *et al*. 2018) pour l'acquisition des termes concernant le domaine médical. L'importance d'une résolution efficace de cette tâche se reflète enfin dans de nombreux domaines de recherche. L'extraction automatique de termes permet d'effectuer des tâches liées à la recherche d'information (comme le repérage de documents pertinents pour une requête donnée), à la fouille de textes (*text mining*), à la construction de ressources terminologiques, etc.

Le point de départ de tous ces travaux concerne l'identification des termes candidats et, par conséquent, le filtrage entre, d'une part, les mots d'ordre général et d'autre part, les termes spécifiques d'un domaine donné. En effet, toutes les études précédemment citées portent (plus ou moins explicitement) sur le concept de « poids » des termes dans une collection des documents afin d'indiquer les différents degrés de pertinence à un domaine. Ce concept a été exprimé, au fil du temps, à travers différentes dénominations. En 1972, Karen Spärck Jones (Sparck Jones 1972) définissait la notion de « spécificité » des termes, comme une valeur calculable en fonction de la fréquence d'apparition des termes dans une collection des documents :

> "[…] the specificity of an individual term is the level of detail at which a given concept is represented."

> "[…] terms should be weighted according to collection frequency, so that matches on less frequent, more specific, terms are of greater value than matches on frequent terms."

En 1996, Kageura et Umino (Kageura et Umino 1996) introduisaient le concept de « termhood » [termicité ou termitude (Humbley 2016)] afin d'indiquer le degré de relation d'une unité linguistique à des concepts spécifiques pour un domaine, une valeur qui peut également être calculée à l'aide d'approches statistiques. Dans ce sens, le degré de termicité d'un terme est donc une valeur déterminante pour la tâche d'extraction automatique des termes et

1 Pour une synthèse de principaux critères, voir le tableau de (Roche 2018).

repose, en général, sur la fréquence d'apparition d'un terme candidat dans le corpus analysé.

Dans ce contexte, nous proposons une perspective différente du concept de « poids » d'un terme technique en nous concentrant sur la notion de « technicité » comme propriété sémantique intrinsèque au terme lui-même. Après sa description théorique, nous procédons à la formalisation de ce concept au moyen d'une fonction qui pondère le degré d'association d'un terme avec un domaine d'intérêt spécifique. En ce sens, notre objectif n'est pas de fournir une nouvelle méthode pour l'extraction automatique des termes, mais plutôt de définir un nouveau paramètre pour leur évaluation.

Cet article est donc organisé comme suit : dans la section 2, nous définissons le concept de « technicité » et nous formalisons cette propriété au moyen de la formule *v-tech*. Dans la section 3, nous présentons une analyse expérimentale menée afin i) de calculer à la fois le degré de technicité et la valeur de termicité de termes médicaux en langue anglaise et ii) d'évaluer la mesure *v-tech* pour la tâche de repérage de documents, en particulier, pour les revues systématiques dans le domaine médical. Enfin, dans la section 4, nous tirons nos conclusions et décrivons nos perspectives.

2. La valeur de technicité d'un terme : le *v-tech*

La raison de ce travail découle de l'analyse des méthodes actuelles pour l'extraction automatique de termes et de la définition implicite de poids d'un terme comme valeur qui dépend de la fréquence et de la distribution de ses occurrences dans un document et/ou un ensemble de documents. Pour cette raison, ces méthodes sont strictement dépendantes du corpus, et le poids d'un terme peut varier en fonction de la collection analysée.

Dans cette étude, nous proposons plutôt de considérer le poids d'un terme comme une valeur qui correspond au niveau de technicité d'un terme pour un domaine donné. La « technicité » est donc une propriété intrinsèque au terme lui-même et pas une valeur statistique dépendant du corpus. Tout en définissant cette propriété, nous ne faisons pas référence aux concepts précédents de 1) « spécificité », comme le niveau de détail auquel un concept donné est représenté, ou 2) « termicité », comme la propriété d'être ou non un terme. Nous ne faisons pas non plus référence à la technicité dans sa connotation négative de 3) « difficulté » de compréhension : à cet égard, de nombreuses études se concentrent, par exemple, sur la terminologie utilisée dans le dialogue patient-médecin et sur les problèmes connexes en matière de com-

préhension et de lisibilité (Tran *et al*. 2009, Bouamor *et al*. 2018, Grabar et Hamon 2016, Ley 1988, Vecchiato et Gerolimich 2013).

La valeur de « technicité » que nous proposons ici dépend du degré d'association du terme à un domaine d'intérêt. Pour fournir une explication intuitive, nous considérons les termes « quadrantectomie » et « patient » qui relèvent du domaine médical. Les deux termes semblent avoir un poids et un degré de technicité différent : le terme « quadrantectomie »[2] désigne un concept unique dans le domaine de la chirurgie et il est répandu chez les spécialistes de ce seul domaine, alors que le terme « patient »[3] véhicule des significations différentes dans plusieurs domaines (médecine, philosophie et linguistique) et son usage est répandu dans un plus grand nombre de domaines de spécialité, et même dans le langage courant. Si nous excluons de cette analyse la diffusion, en tant que concept proche de la fréquence d'apparition d'un terme dans un corpus réel ou imaginaire, nous nous concentrons sur la définition de « technicité » comme propriété dont la valeur est inversement proportionnelle au nombre de domaines dans lesquels un terme apparaît. En ce sens, plus un terme est polysémique, c'est-à-dire plus il relève de plusieurs domaines, moins il sera technique, car il ne sera ni monoréférentiel ni exclusif. Une caractéristique souhaitable de la terminologie employée pour un domaine est celle d'être monoréférentielle et spécifique (Guilbert 1973) : la relation entre signe et référent devrait être univoque afin d'éviter les ambigüités et la polyvalence du point de vue sémantique. Or, un signe linguistique (un terme) qui désigne plus de référents perd en « technicité », car il n'est plus exclusif d'un seul domaine. À partir de cette définition, nous avons formalisé ce principe en attribuant une valeur numérique à la technicité d'un terme. Nous introduisons donc la valeur que nous appellerons *v-tech* par la formule suivante :

$$\text{v-tech (t)} = \begin{cases} e^{-\lambda d_t}, & d_t > 0 \\ 0, & d_t = ? \end{cases}$$

où *t* est le terme pour lequel la valeur de technicité (*v-tech*) est calculée, où d_t est le nombre de domaines dans lesquels un terme *t* apparaît, et où l est un paramètre qui adapte la rapidité avec laquelle le terme *t* perd en technicité a1 mesure qu'augmente le nombre de domaines qui l'adoptent. Le principe à la base de *v-tech* est que les termes polysémiques, ayant plusieurs domaines d'utilisation, auront une valeur qui tendra vers 0. Les termes n'ayant pas un

2 http://www.granddictionnaire.com/ficheOqlf.aspx?Id_Fiche=26527442
3 http://www.cnrtl.fr/definition/patient

domaine clairement explicité dans une ressource auront une valeur de *v-tech* égale à 0.

L'image suivante (FIG. 1) illustre l'évolution de *v-tech* en fonction de d_t lorsque le paramètre l varie :

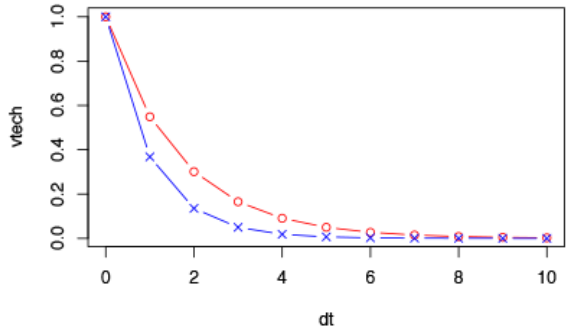


FIG. 1 – *Valeurs de v-tech pour l=0,6 (ligne rouge) et l=1 (ligne bleu)*

Or, pour calculer le score *v-tech*, il est nécessaire de s'appuyer sur des ressources linguistiques montrant tous les domaines associés à une unité linguistique. Par conséquent, le poids d'un terme n'est plus une valeur qui dépend du corpus et calculée en fonction de la fréquence de ses occurrences, mais il devient une valeur dépendant des ressources et basée sur l'exhaustivité des données relatives aux domaines dans lesquels le terme est utilisé.

À notre connaissance, BabelNet est actuellement la ressource la plus complète et la plus structurée qui contient les domaines associés aux termes (Camacho-Collados et Navigli 2017). Cependant, il y a d'autres ressources qui, pour certains termes techniques, collectent plus de domaines que BabelNet. Pour cette raison, afin de calculer la valeur *v-tech* de manière complète et précise, nous avons mené les expériences décrites dans la section suivante en rassemblant les informations fournies par BabelNet, Termium Plus,[4] la base

4 https://www.btb.termiumplus.gc.ca

de données IATE,[5] le dictionnaire en ligne Merriam-Webster[6] et le Grand Dictionnaire Terminologique.[7]

3. Expériences

Dans cette section, nous présentons une analyse expérimentale menée dans le but de : 1) donner un « poids » à un ensemble de termes en comparant la valeur de *v-tech*, telle que définie ci-dessus, à d'autres mesures de termicité basées sur la collection, 2) évaluer l'intégration de la formule *v-tech* dans le score d'un modèle de repérage de documents pour la tâche de revue systématique. Une revue systématique consiste à identifier et à collecter toutes les études, publiées ou non, traitant d'un sujet donné.[8]

Pour cette analyse, nous avons choisi le domaine médical et nous avons utilisé la collection de documents en langue anglaise fournis par le colloque CLEF 2018 (*Conference and Labs of the Evaluation Forum*),[9] pour l'accomplissement de la tâche nommée « *e-Health Technology Assisted Reviews in Empirical Medicine* » (Kanoulas *et al*. 2018). L'ensemble de données comprend : i) 30 sujets médicaux, à savoir les besoins d'informations médicales à satisfaire fournis par un médecin lors de l'accomplissement de revues systématiques ; ii) un ensemble de documents de PubMed[10] ; et iii) un ensemble de jugements de pertinence.

3.1. Le paramètre *V-tech* confronté à d'autres mesures de termicité

Dans la première partie de notre analyse, nous nous concentrons sur l'attribution d'un score représentant le poids d'un ensemble de termes médicaux. Notre but est de comparer la formule *v-tech* avec d'autres mesures de termicité basées sur le corpus. À cet égard, nous procédons d'abord à l'extraction manuelle des termes médicaux identifiés dans l'ensemble des 30 sujets de la collection CLEF 2018. Ensuite, pour chaque terme, nous avons effectué une analyse qualitative en vérifiant tous les domaines d'utilisation de ces termes sur les ressources mentionnées ci-dessus. Nous avons collecté un total de 192 termes ayant un comportement sémantique différent : 104 termes

5 https://iate.europa.eu/home
6 https://www.merriam-webster.com
7 http://www.granddictionnaire.com
8 https://ccf.cochrane.org/revues-cochrane
9 http://clef2018.clef-initiative.eu/index.php
10 https://www.ncbi.nlm.nih.gov/pubmed

monosémiques apparaissent dans un seul domaine (médecine) ou sous-domaine (oncologie, pathologie, chirurgie) ; les 88 autres termes polysémiques envisagent jusqu'à 13 domaines d'utilisation. Dans le TAB.1, nous pouvons observer quelques résultats de notre analyse qualitative : des termes comme « cytology » et « radiculopathy », en tant que monosémiques, ont une valeur de *v-tech* élevée, alors que d'autres termes comme « screening » et « marker » qui sont bien attestés dans plusieurs domaines, ont plutôt une valeur de *v-tech* très basse.

Ensuite, nous avons calculé la fréquence d'occurrence TF (*Term Frequency*) de ces termes dans la collection de documents de PubMed, la fréquence de documents DF (*Document Frequency*) dans lesquels apparaissent les termes analysés et le C-Value visant à donner un poids aux termes complexes (par exemple « globule rouge » ou « douleur musculaire »). Dans le TAB. 2, nous montrons un exemple de trois termes pour deux sujets (A et B correspondent respectivement aux sujets CD010680 et CD008892 de la collection CLEF 2018) avec leur TF, DF, C-Value et la valeur *v-tech* calculée avec l=0.6. Si l'on considère ces deux sous-ensembles de documents, les trois termes ont des fréquences variables et la valeur de leur termicité (TF, DF, C-Value) change significativement en fonction des documents analysés. Si l'on confronte les trois mesures de termicité basées sur le corpus, le terme « pulmonary », par exemple, a un poids complètement différent dans les deux collections de documents A et B. D'autre part, le poids de ces termes basés sur la valeur de *v-tech* reste exactement le même, dans la mesure où la technicité est calculée sur les domaines d'utilisation et définie sur la ressource linguistique.

Terme	Domaine	Ressource
Diagnostic	Information Technology, Medicine, Systems Analysis, Meteorological Forecasting, Psychology, Servicing and Maintenance, Law	BabelNet, Termium plus, IATE
Endemic	Biology, Epidemiology	BabelNet, Termium plus
Cytology	Biology	BabelNet, Merriam-Webster

Terme	Domaine	Ressource
Screening	Economics, Security, Air Transport, Water Treatment, Records Managment, Football, Epidemiology, Cinematography, Press, Waste Managment, Law, Medicine	BabelNet, Termium plus
Marker	Aeronautical, agriculture, army, art, biology, electronic, geology, linguistic, medicine, information science, sociology, sport, telecommunication	BabelNet, Le grand dictionnaire terminologique
Radiculopathy	Pathology	BabelNet, Merriam-Webster

Tab. 1 – *Liste partielle de termes médicaux, domaines d'utilisation et ressources employées.*

Topic	Term	TF	DF	C-Value	v-tech
A	Pulmonary	994	303	993	0.549
A	Typhoid	1	1	0	0.549
A	Resonance	392	273	391	0.015
B	Pulmonary	8	5	6.86	0.549
B	Typhoid	29	13	27.5	0.549
B	Resonance	8	6	6.85	0.015

Tab. 2 – *Comparaison entre TF, DF, C-Value et V-tech de trois termes dans deux sujets.*

3.2. Revues systématiques médicales avec *V-tech*

Dans cette deuxième expérience, nous visons à évaluer l'impact de la valeur *v-tech* pour la tâche de revues systématiques dans le domaine médical, c'est-à-dire pour le repérage de tous les documents pertinents à un sujet médical donné. Pour accomplir cette tâche, de nombreux modèles de repérage permettent d'attribuer un poids aux termes qui constituent la requête : voir le BM25 (Robertson et Zaragoza 2009). Notre hypothèse est qu'en combinant le poids donné par le modèle de base et le poids calculé à partir de la fonction

v-tech, la précision des modèles de repérage s'améliore en termes de documents pertinents récupérés.

Comme modèle de base, nous utilisons la variante CAL (*Continuous Active Learning)* (Di Nunzio 2018) du modèle de repérage BM25. En particulier, suivant l'approche suggérée par (Ventura 2014) avec un document *d* et une requête *q* le nouveau score BM25 est :

$$score(d,q) = \sum_{t \in q \cap d} w_t^{BM25}(1 + \text{v-tech(t)})$$

où, pour chaque terme de la requête qui apparaît dans le document, nous multiplions le poids du BM25 du terme w_t^{BM25} par sa valeur *v-tech*. Nous ajoutons 1 à la valeur *v-tech* afin de prendre en compte les termes pour lesquels le nombre de domaines est inconnu et qui ont, par définition, une valeur de *v-tech* égale à 0.

Dans l'image suivante (FIG. 2), nous présentons les résultats obtenus, en termes de repérage sur la moyenne des 30 sujets, de l'application du modèle de base (baseline) et du modèle de base combiné avec la valeur *v-tech* (v-tech l=0.6). Les mesures d'évaluation sont : 1) la précision à *k* de documents extraits (P@k, c'est-à-dire le rapport entre le nombre de documents pertinents dans les premiers *k* documents et la variable *k*), et 2) le rappel à *R* documents pertinents (c'est-à-dire le nombre de documents pertinents repérés divisé par le nombre *R* total de documents pertinents présents dans la collection).

Les meilleurs résultats (en gras) confirment que la combinaison de la valeur *v-tech* dans le score de repérage du BM25 augmente la précision des documents pertinents récupérés dans les premiers 10, 20, et 50 documents. En outre, le rappel global améliore par rapport au modèle de base.

model	# docs	# rel docs	# rel ret	P@10	P@20	P@50	Recall
baseline	217507	3964	886	0.340	0.345	0.348	0.447
vtech $\lambda = 0.6$	217507	3964	**922**	**0.353**	**0.358**	**0.354**	**0.460**

FIG. 2 – *Résultats moyens sur 30 sujets obtenus en utilisant le modèle de base BM25 et le modèle combiné avec le v-tech.*

4. Conclusions et perspectives

Cette étude repense le concept de poids d'un terme en se concentrant sur la notion de « technicité » comme propriété intrinsèque des termes. L'idée de base est que la valeur de technicité d'un terme est inversement proportionnelle à sa nature polysémique. Sur la base de ce principe, nous avons développé une définition formelle à travers la formule *v-tech*. Dans les expériences menées, nous avons montré que le poids donné par le *v-tech* est une valeur différente d'autres mesures statistiques basées sur un corpus (telles que TF-IDF, C-value, etc.), puisqu'elle est calculée à partir des informations fournies par les ressources linguistiques sur les domaines d'utilisation. De plus, l'intégration de la valeur *v-tech* dans le modèle de base du BM25 a montré que les performances de repérage s'améliorent en moyenne dans des tâches spécifiques telles que la récupération de documents médicaux pour les revues systématiques.

À partir de cette première analyse, nous avons constaté la nécessité de créer une ressource terminologique structurée qui puisse intégrer complètement les informations en ligne. Dans les expériences menées jusqu'à présent, nous avons rassemblé les informations fournies par des ressources différentes. En outre, nous nous proposons de travailler sur un corpus de textes en langue française afin de comparer le différent degré de technicité des termes, telle que celle-ci a été définie, dans une perspective multilingue. Enfin, spécifiquement pour le domaine médical, nous visons à approfondir la relation entre technicité et difficulté de compréhension dans le contexte du dialogue médecin-patient.

Remerciements

Je tiens à remercier le Professeur Giorgio Maria Di Nunzio du Département d'Ingénierie de l'Information (Université de Padoue) pour son aide dans la définition formelle de *v-tech* et pour les expériences menées jusqu'à présent.

Références

Amjadian, Ehsan, Diana Inkpen, T Sima Paribakht, and Farahnaz Faez. 2018. "Distributed Specificity for Automatic Terminology Extraction." *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 24 (1): 23-40.

Bouamor, Dhouha, Leonardo Campillos Llanos, Anne-Laure Ligozat, Sophie Rosset, and Pierre Zweigenbaum. 2016. "Transfer-Based Learning-to-Rank Assessment of Medical Term Technicality." In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2312-2316.

Bourigault, Didier, Christian Jacquemin, and Marie-Claude L'Homme. 2001. *Recent Advances in Computational Terminology*. Vol. 2. John Benjamins Publishing.

Camacho-Collados, Jose, and Roberto Navigli. 2017. "BabelDomains : Large-Scale Domain Labeling of Lexical Resources." In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 2, Short Papers*, 223-228.

Church, Kenneth, William Gale, Patrick Hanks, and Donald Hindle. 1991. "Using Statistics in Lexical Analysis." *Lexical Acquisition : Exploiting on-Line Resources to Build a Lexicon* 115 : 164.

Church, Kenneth Ward, and Patrick Hanks. 1990. "Word Association Norms, Mutual Information, and Lexicography." *Computational Linguistics* 16 (1) : 22-29.

Di Nunzio, Giorgio Maria. 2018. "A Study of an Automatic Stopping Strategy for Technologically Assisted Medical Reviews." In *European Conference on Information Retrieval*, 672-677. Springer.

Drouin, Patrick. 2003. "Term Extraction Using Non-Technical Corpora as a Point of Leverage." *Terminology* 9 (1) : 99-115.

Drouin, Patrick, Natalia Grabar, Thierry Hamon, Kyo Kageura, and Koichi Takeuchi. 2018. "Computational Terminology and Filtering of Terminological Information." *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 24 (1) : 1-6.

Foo, Jody. 2012. "Computational Terminology : Exploring Bilingual and Monolingual Term Extraction." PhD Thesis, Linköping University Electronic Press.

Frantzi, Katerina T, Sophia Ananiadou, and Junichi Tsujii. 1998. "The C-Value/Nc-Value Method of Automatic Recognition for Multi-Word Terms." In *International Conference on Theory and Practice of Digital Libraries*, 585-604. Springer.

Grabar, Natalia, and Thierry Hamon. 2016. "A Large Rated Lexicon with French Medical Words." In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2643-2648. Portorož, Slovenia : European Language Resources Association (ELRA).

Guilbert, Louis. 1973. "La Spécificité Du Terme Scientifique et Technique." *Langue Française*, no. 17 : 5-17.

Humbley, John. 2016. "Catherine Resche (Dir.), Terminologie et Domaines Spécialisés, Approches Plurielles. Paris : Classiques Garnier, Rencontres 143, Série Linguistique 2, 2015." *ASp. La Revue Du GERAS*, no. 70 : 127-132.

Inkpen, Diana, T Sima Paribakht, Farahnaz Faez, and Ehsan Amjadian. 2016. "Term Evaluator : A Tool for Terminology Annotation and Evaluation." *International Journal of Computational Linguistics and Applications* 7 (2).

Jones, Karen Spärck. 1972. "A Statistical Interpretation of Term Specificity and Its Application in Retrieval." *Journal of Documentation*.

Kageura, Kyo, and Bin Umino. 1996. "Methods of Automatic Term Recognition : A Review." *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 3 (2) : 259-289.

Kanoulas, Evangelos, Rene Spijker, Dan Li, and Leif Azzopardi. 2018. "CLEF 2018 Technology Assisted Reviews in Empirical Medicine Overview." In *CLEF 2018 Evaluation Labs and Workshop : Online Working Notes. CEUR-WS, France*, 1-20.

Ley, Philip. 1988. *Communicating with Patients : Improving Communication, Satisfaction and Compliance.* Croom Helm.

Nakagawa, Hiroshi. 2001. "Experimental Evaluation of Ranking and Selection Methods in Term Extraction." *Bourigault D, L'Homme M.-C., Jacquemin C.(Éd.), Recent Advances in Computational Terminology, John Benjamins Publishing Company*, 303-26.

Planas, Emmanuel. 2012. "BiTermEx Un Prototype d'extraction de Mots Composés à Partir de Documents Comparables via La Méthode Compositionnelle (BiTermEx, A Prototype for the Extraction of Multiword Terms from Comparable Documents through the Compositional Approach) [in French]." In Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, Volume 2 : TALN, 415-422. Grenoble, France : ATALA/AFCP.

Robertson, Stephen, Hugo Zaragoza, and others. 2009. "The Probabilistic Relevance Framework : BM25 and Beyond." *Foundations and Trends® in Information Retrieval* 3 (4) : 333-389.

Roche Mathieu. 2018. "Définition pluridisciplinaire de la notion de "terme"". In : TOTh 2017. Terminologie et ontologie : Théories et applications. Roche Christophe (ed.). Chambéry : Université Savoie Mont Blanc, 63-72.

Salton, Gerard, and Chung-Shu Yang. 1973. “On the Specification of Term Values in Automatic Indexing.” *Journal of Documentation* 29 (4) : 351-372.

Sandoval, Antonio Moreno, Julia Díaz, Leonardo Campillos Llanos, and Teófilo Redondo. 2019. “Biomedical Term Extraction : NLP Techniques in Computational Medicine.” *IJIMAI* 5 (4) : 51-59.

Simon, Nisha Ingrid, and Vlado Kešelj. 2018. “Automatic Term Extraction in Technical Domain Using Part-of-Speech and Common-Word Features.” In *Proceedings of the ACM Symposium on Document Engineering 2018*, 51. ACM.

Tran, Thi Mai, H Chekroud, P Thiery, and A Julienne. 2009. “Internet et Soins : Un Tiers Invisible Dans La Relation Médecine/Patient.” *Ethica Clinica* 53 : 34-43.

Vecchiato, Sara, and Sonia Gerolimich. 2013. “La Langue Médicale Est-Elle « trop Complexe » ?” *Nouvelles Perspectives En Sciences Sociales : Revue Internationale de Systémique Complexe et d’études Relationnelles* 9 (1) : 81-122.

Ventura, Juan Antonio Lossio, Clement Jonquet, Mathieu Roche, and Maguelonne Teisseire. 2014. “Towards a Mixed Approach to Extract Biomedical Terms from Text Corpus.” *International Journal of Knowledge Discovery in Bioinformatics (IJKDB)* 4 (1) : 1-15.

Vu, Thuy, Aiti Aw, and Min Zhang. 2008. “Term Extraction through Unithood and Termhood Unification.” In *Proceedings of the Third International Joint Conference on Natural Language Processing : Volume-II*.

Abstract

In this study, we propose a different perspective on the concept of weight of technical terms by focusing on the notion of “technicality” as a semantic property of the linguistic unit itself. The basic idea is that the value of technicality of a term is inversely proportional to its polysemic nature. We formalise the v-tech formula and carry out an experimental evaluation in order to 1) compare the v-tech value with other collection-based measures of termhood, and 2) integrate the v-tech formula in the score of a retrieval model for systematic reviews task for the medical domain.

Gibran 2.0 : analyse morphosyntaxique de l'arabe par une approche linguistique

Youcef Ihab Morsi, Iana Atanassova

CRIT, Université de Bourgogne Franche-Comté
30 rue Mégevand, 25000 Besançon, France
youcef.morsi@univ-fcomte.fr, iana.atanassova@univ-fcomte.fr

Résumé. Nous proposons un outil d'analyse morphosyntaxique de l'arabe standard moderne qui est basé sur l'étude linguistique des schèmes. Notre premier objectif est de proposer des règles linguistiques pour l'analyse morphosyntaxique et évaluer notre méthode sur le corpus annoté Arabic-PADT UD treebank. Ce dernier, destiné au développement d'outils par apprentissage automatique, comporte de nombreuses occurrences de mots annotés avec une étiquette X, signalant une classe inconnue. Le deuxième objectif de notre étude est de proposer une méthode pour améliorer cette ressource en annotant ces occurrences inconnues avec les classes adéquates. Ainsi notre méthode vise à optimiser ce corpus annoté. Les deux évaluations montrent une F-mesure de 0,92 et 0,90 pour les analyses morphosyntaxiques.

1. Introduction

Nous proposons l'outil d'analyse morphosyntaxique de l'arabe Gibran 2.0 qui est basé sur l'étude linguistique des schèmes (Blachère & Gaudefroy-Demombynes, 1966). Notre objectif est de proposer des règles linguistiques pour l'analyse morphosyntaxique et évaluer notre méthode sur le corpus annoté Arabic-PADT UD treebank. En effet, ce corpus qui est destiné au développement d'outils par apprentissage automatique comporte de nombreuses occurrences de mots annotés avec une étiquette X, signalant une classe inconnue pour une raison ou une autre. Le deuxième objectif de notre étude est de proposer une méthode pour améliorer cette ressource en annotant ces occurrences avec les classes adéquates. Le lexème arabe a la particularité d'avoir une base consonantique trilitère ou quadrilatère (Blachère & Gaudefroy-Demombynes, 1966) qui accepte des combinaisons (consonnes/

voyelles-longues) pour former des structures internes. Celles-ci définissent par exemple la déclinaison des lexèmes en verbes, adjectifs, noms, etc. Les mots et groupes de mots s'agglutinent pour générer des formes complexes (Pasha *et al*., 2014), exemple : la forme أتتذكروننا correspond en français à *Est-ce que vous vous souvenez de nous ?* (Belguith & Chaaben, 2006). Ce phénomène d'agglutination nous incite à analyser la langue arabe différemment que les langues latines, voire à oblitérer certaines frontières entre la morphologie et la syntaxe.

> *Arabic has a high degree of ambiguity resulting from its diacritic-optional writing system and common deviation from spelling standard* (Aloulou, 2003)

Les travaux de (Debili *et al*., 2002) ont montré que l'absence de vocalisation dans les textes arabes augmente le niveau d'ambiguïté morphosyntaxique des lexèmes de 66 % à 95 % pour les textes. Les signes diacritiques sont le plus souvent utilisés pour l'apprentissage de la langue mais sont de moins en moins utilisés dans l'arabe standard (journaux, médias, réseaux sociaux, etc.).

Dans le cadre des travaux sur l'analyse morphosyntaxique de la langue arabe, nous pouvons citer les travaux de (Kammoun *et al*., 2010) avec l'analyseur MORPH2 (Kammoun *et al*., 2010). Ce dernier se base sur une analyse linguistique opérant sur le concept de racine, ajoutée à une liste de caractéristiques morphologiques qui permettent d'identifier les différentes catégories et ainsi désambiguïser des phénomènes liés à l'agglutination, à la non vocalisation, etc. MORPH2 fait appel à un lexique conséquent de racines et d'un lexique des formes canoniques pour la classe des noms ambigus.

Les travaux de (Pasha *et al*., 2014) sur l'outil MADAMIRA proposent une analyse et une désambiguïsation morphologique en arabe, pour extraire des informations sous forme de listes relatives aux caractéristiques d'un lexème (POS tag, genre, lemme, etc.). Celle-ci se base sur un ensemble de modèles d'apprentissage tel que les SVM et les N-grams pour prédire les formes et les caractéristiques d'un mot dans son contexte. Une fois l'analyse faite, les résultats sont classés par pondération pour chaque mot et le meilleur score est retenu pour définir la classe finale. Cet outil présente des résultats de 84.1 % de précision.

Les travaux de (Ghoul, 2013) consistent à développer des ressources destinées aux applications d'apprentissage automatique. Il propose un corpus d'entraînement à l'outil TreeTagger et fournit un étiquetage morphosyntaxique qui se base sur l'apprentissage par arbres de décision. Ce modèle probabiliste per-

met de constituer des couples de mot / catégorie (Ghoul, 2013). Les résultats donnés par cette recherche stipulent que la précision de cet outil est de 86 %.

Alkhalil Morpho Sys 2 (Boudchiche & Mazroui, 2016) est un analyseur morphologique qui attribue pour chaque mot un lemme unique en prenant en compte le contexte des mots analysés. Cet outil analyse dans un premier temps les lexèmes et leur attribue plusieurs étiquettes possibles. Ensuite, une phase de désambiguïsation intervient en contexte prenant comme base les modèles de Markov cachés, les techniques de lissage et l'algorithme de Viterbi.

2. Méthodologie proposée

Dans cette section nous décrivons la démarche méthodologique pour la création de notre système d'analyse morphologique de la langue arabe. Notre analyseur morphologique comprend plusieurs traitements successifs qui permettent de désambiguïser les étiquettes possibles pour les formes en arabe :

1. Couche morphologique, qui s'appuie sur les schèmes arabes comme un trait distinctif et propose une ou plusieurs étiquettes possibles pour chaque mot dans le texte ;
2. Couche syntaxique, qui fait appel à des règles contextuelles dans les cas où plusieurs étiquettes sont possibles. Ces règles ont pour objectif de désambiguïser la catégorie grammaticale quand le contexte du mot (formes et étiquettes se trouvant à gauche et à droite du mot dans la même phrase) le permet.

2.1. Couche morphologique

La morphologie arabe est soumise à des règles strictes obéissant à des patrons dérivationnels. C'est sur la base de ces considérations que nous avons élaboré la couche morphologique. Le concept de schème, appelé thème chez (Blachère & Gaudefroy-Demombynes, 1966 ; Cohen), s'applique à la racine et intervient lors de la déclinaison, dérivation, conjugaison, en définissant les possibilités d'affixation. Ce patron phonologique est une combinaison de consonnes et voyelles, par exemple : CCVC, CCV, CCC, etc.

Schème	Exemple	Traduction	Règle	
فاعل	شارك	participer	(CONSONNE + S-VOYELLE + CONSONNE(2))	R1
تفاعل	تنازع	lutter	(ت + CONSONNE + S-VOYELLE + CONSONNE(2))	R2
افتعل	اكتسب	acquérir	(ا + CONSONNE + ت + CONSONNE(2))	R3
استفعل	استقبل	accueillir	(است + CONSONNE(3))	R4
استفعال	استقبال	accueil	(است + CONSONNE(2) + ا + CONSONNE)	R5

Tab. 1 – Règles morphologiques

Le tableau 1 présente un échantillon des règles morphologiques que nous avons créées. La conjugaison du verbe « participer » R1 est dirigée par la combinaison CVCC. Pour le verbe « lutter » R2, cette combinaison est la même, mais avec l'ajout d'un préfixe |t| ت. Pour le verbe « acquérir » R3 nous avons l'ajout d'un préfixe |a| ا et d'un infixe |t| ت à la position [2]. Enfin dans l'exemple R4 « accueillir », le verbe admet un préfixe |ist| است avec la combinaison CCC.

Tous ces exemples appartiennent à un schème verbal à racine trilitère. Les lettres de cette racine peuvent être toutes les consonnes ou voyelles longues de la langue arabe. À cette base s'ajoute l'ensemble des préfixes et suffixes relatifs aux personnes, temps et modes du verbe. Chaque règle définit de ce fait une direction dérivationnelle, qui est une sorte de prédiction des formes possibles à travers l'ajout, la modification ou la suppression d'une ou plusieurs lettres, etc. Ce même principe s'applique aussi aux classes des noms et des adjectifs subissant des schèmes qui régissent leurs flexions. L'ajout de la voyelle longue ا|a :| dans l'exemple R5 du tableau 1 génère un lexème de type nom R5 et représente ainsi une nouvelle règle de déclinaison du nom.

Concernant les particules dans la langue arabe (prépositions, conjonctions, pronoms, etc.), nous avons fait le choix de constituer une liste finie contenant toutes les formes de ces lexèmes. Cette liste n'est pas exhaustive et peut être enrichie. Ces particules acceptent aussi des préfixes et suffixes selon leurs caractéristiques. Les pronoms relatifs, par exemple, s'agglutinent à des conjonctions de coordinations. Cette classe fermée possède aussi des règles de bonne formation en termes de préfixation et suffixation.

Pour le traitement des noms propres, qui est un problème important en arabe dû à l'absence de capitalisation, nous avons utilisé le corpus annoté d'entités nommées ANER (Benajiba *et al*., 2007). Ce corpus possède plus

de 150 000 occurrences, dont nous avons extrait tous les noms de personne, noms de lieu et noms d'organisation. La liste a été complétée manuellement avec d'autres noms propres (de personnes, de villes et de capitales) extraits à partir de Wikipédia, ainsi que des noms de mesures, de jours et de mois. La liste finale contient 33 065 occurrences.

2.2. Couche syntaxique

Les règles contextuelles nous permettent d'introduire des contraintes syntaxiques afin de lever l'ambiguïté de certaines formes. Nous avons élaboré plus de 1 140 règles contextuelles.

ID	: Conditions	Résultat	
rc1	: "الى" + NOUN/ADJ	1 NOUN	45
rc5	: "الّذي" + NOUN/ADJ/VERB	1 VERB	46
rc48	: DET + NOUN/ADJ/VERB	1 NOUN	47

Tab. 2 – Règles contextuelles avec deux éléments.

Le tableau 2 présente des exemples de règles contextuelles qui se basent sur le contexte droit et gauche des mots analysés. La règle 45 indique que lorsque nous avons plusieurs étiquettes pour un mot qui est dans le contexte rc1, l'étiquette de ce mot prend la valeur NOUN. La règle 46 indique que seuls les verbes peuvent être acceptés dans le contexte rc5. La règle 47 indique qu'après un déterminant la classe qui est possible pour le mot suivant est NOUN.

3. Résultats

Pour évaluer notre outil, nommé Gibran 2.0, nous nous sommes basés sur le corpus annoté Arabic-PADT UD treebank (Smrz *et al*., 2002 ; Hajic *et al*., 2004 ; Smrz *et al*., 2007) qui est disponible en ligne et en accès libre. Arabic-PADT UD treebank est un corpus annoté semi automatiquement issu du projet Universal Dependencies[1]. L'objet principal de ce projet est de développer des annotations dites universelles sur des corpus multilingues afin de promouvoir et faciliter le développement d'analyseurs morphosyntaxiques pour des diffé-

1 *http://universaldependencies.org*

rentes langues. Le corpus en arabe contient plus de 282 380 occurrences. Il est annoté avec un ensemble d'étiquettes morphosyntaxiques. Certaines formes dans ce corpus n'ont pas pu être annotées et portent l'étiquette X. Ces formes constituent environ 7,89 % du corpus, soit 22 298 formes.

Pour cette évaluation notre objectif est double : évaluer notre méthode morphosyntaxique et comparer les résultats au corpus annoté Arabic-PADT UD treebank ; et proposer des annotations pour les formes non annotées dans ce corpus.

Type 0	Résultats vides pour Gibran 2.0 et Arabic-PADT UD treebank.
Type 1	Résultat contradictoire.
Type 2	Gibran 2.0 retourne un résultat vide et Arabic-PADT UD treebank donne un résultat.
Type 3	Arabic-PADT UD treebank donne un résultat et Gibran 2.0 plusieurs.
Type 4	Gibran 2.0 et Arabic-PADT UD treebank donnent des résultats équivalents et pas vides.
Type 5	Arabic-PADT UD treebank donne un résultat vide et Gibran 2.0 renvoie plusieurs résultats.
Type 6	Arabic-PADT UD treebank donne un résultat vide et Gibran 2.0 renvoie un seul résultat.

Tab. 3 – Typologie des résultats.

Dans le tableau 3, nous avons établi 7 types de résultats (de 0 à 6) dans l'objectif de pouvoir comparer les résultats de notre analyse aux annotations du corpus. Le tableau 4 montre les différentes étiquettes qui ont été attribuées par notre système aux occurrences étiquetées (étiquette X exclue).

	Evaluation stricte		Evaluation large	
	Précision	Rappel	Précision	Rappel
NOUN	0,80	0,87	0,81	0,92
VERB	0,42	0,46	0,87	0,92
ADJ	0,44	0,43	0,83	0,73
ADP	0,98	0,91	0,98	0,91
CONJ	1	1	1	1
PART	0,96	0,85	0,96	0,85
DET	0,92	0,88	0,92	0,88
NUM	1	0,99	1	0,99
PRON	0,89	0,86	0,89	0,86
PUNCT	0,99	0,99	0,99	0,99
SYM	1	1	1	1
ADV	1	1	1	1
F-mesure	0,85		0,92	

TAB. 4 – Évaluation précision, rappel et F-mesure.

Nous avons calculé la précision, le rappel et la F-mesure pour chaque étiquette présente dans le corpus et avons établi deux types d'évaluation : une évaluation stricte qui prend en compte seulement les résultats de type 4 (les résultats correctement étiquetés et ayant une seule proposition), et une évaluation large qui prend en compte l'ensemble des résultats de types 3 et 4. L'évaluation stricte montre un résultat très bas pour les classes VERB et ADJ par rapport aux autres classes comme NOUN et ADP.

Arabic-PADT UD treebank	الاصلاح الزراعي	la réforme agricole	NOUN + ADJ	1
Gibran 2.0	الاصلاح الزراعي	la réforme agricole	NOUN + ADJ ‖ NOUN	2
Arabic-PADT UD treebank	صحيفتي	journaux (au duel)	NOUN	3
Gibran 2.0	صحيفتي	journaux (au duel)	NOUN ‖ ADJ	4
Arabic-PADT UD treebank	يهدف	vise	VERB	5
Gibran 2.0	يهدف	vise	VERB ‖ NOUN ‖ ADJ	6

TAB. 5 – *Résultats de type 3.*

Le tableau 5 présente des exemples de résultats de type 3, où Gibran 2.0 donne plusieurs résultats alors que le corpus Arabic-PADT UD treebank identifie une seule étiquette. L'exemple 2 du tableau 5, agricole, est identifié comme appartenant à deux classes morphologiques (ADJ et NOUN). L'exemple 4, journaux, est étiqueté comme étant un nom au duel et un adjectif. L'exemple 6 indique que l'étiquette correspondante est bien présente mais d'autres étiquettes sont possibles.

Pour l'évaluation large, tableau 4, les résultats deviennent assez uniformes et se rapprochent d'une moyenne de 0,92 en F-mesure. L'évaluation stricte donne une moyenne de 0,85 en F-mesure.

À partir du corpus Arabic-PADT UD treebank, nous avons établi une deuxième évaluation (échantillon de 2044 occurrences dont 300 formes non reconnues ayant l'étiquette X). Ces occurrences ont été annotées manuellement pour pouvoir les comparer avec les résultats de Gibran 2.0.

	Evaluation stricte		Evaluation large	
	Précision	Rappel	Précision	Rappel
ADP	1	1	1	1
NOUN	0,51	0,41	1	0,80
NUM	1	1	1	1
PRON	1	1	1	1
Abbr	1	1	1	1
VERB	0	0	1	1
ADJ	1	1	1	1
Total	0,78	0,77	1	0,82
F-mesure	0,77		0,90	

Tab. 6 – *Évaluation précision, rappel et F-mesure.*

Le tableau 6 montre les différentes étiquettes qui ont été attribuées par notre système aux formes non reconnues. Nous constatons que la F-mesure est de 0.77 pour l'évaluation stricte et 0.90 pour l'évaluation large. L'évaluation stricte montre que la classe des noms est très ambiguë parmi les formes non reconnues, et celle des verbes donne un résultat nul. Ces résultats montrent que notre approche arrive à désambiguïser et améliorer les annotations des formes non reconnues dans le corpus Arabic-PADT UD treebank. Une meilleure performance peut être atteinte en développant la portée de nos règles contextuelles.

4. Conclusion

Nous avons proposé une méthodologie basée sur des règles linguistiques et développé l'outil Gibran 2.0. Le concept de schème est au centre de notre approche. Cette dernière se base sur deux couches, à savoir : morphologique, qui attribue toutes les étiquettes possibles pour chaque occurrence, et syntaxique, qui désambiguïse les étiquettes si nécessaires.

Les difficultés principales dans une telle approche sont de pouvoir délimiter la portée des règles morphologiques. En effet, les classes des noms et adjectifs en arabe ont des propriétés communes qui les rendent ambiguës. Les schèmes donnant lieu essentiellement à des règles de nature générative, une règle morphologique peut couvrir à la fois un schème nominal (nom et adjectif) mais aussi un schème verbal. Une telle approche doit s'appuyer donc sur le contexte syntaxique des occurrences pour procéder à la désambiguïsation.

Nous avons mis en place deux types d'évaluations : une évaluation stricte qui considère comme vrai positif seulement les résultats uniques (une seule étiquette) et une évaluation large qui prend en considération les étiquettes uniques mais aussi les étiquettes multiples (plusieurs étiquettes pour une occurrence). Ces deux évaluations ont porté sur deux expérimentations différentes : la première regroupe tous les résultats des classes annotées (excluant l'étiquette X) et la deuxième porte seulement sur les occurrences inconnues étiquetées X dans Arabic-PADT UD treebank.

Les résultats montrent que la première expérimentation présente une moyenne en F-mesure de 0,85 pour l'évaluation stricte et 0,92 pour l'évaluation large. L'écart entre les deux évaluations soutient notre précédent constat : le contexte syntaxique peut améliorer cette performance. La deuxième expérimentation présente une moyenne en F-mesure de 0,77 pour l'évaluation stricte et 0,90 pour l'évaluation large. En d'autres termes, Gibran 2.0 serait en mesure de compléter les étiquettes X dans Arabic-PADT UD treebank avec une précision de 0,78.

Les corpus annotés ont une place primordiale dans les applications basées sur l'apprentissage automatique. En ce sens, cette recherche propose une solution concrète pour l'optimisation d'un corpus annoté à grande échelle. En outre, l'expérimentation montre qu'une approche qui fait appel à des connaissances linguistiques permet d'obtenir des résultats pour les analyses morpho-syntaxiques dont la qualité est comparable au corpus annoté Arabic-PADT UD treebank qui a été créé dans une perspective d'apprentissage automatique.

Références

Aloulou, Chafik. 2003. « Analyse syntaxique de l'Arabe : Le système MASPAR ». *RECITAL*, Nantes : France.

Belguith, Lamia & Chaaben, Nouha. 2006. « Analyse et désambiguïsation morphologiques de textes arabes non voyellés ». Actes de la 13e conférence sur le Traitement Automatique des Langues Naturelles, pp. 493-501.

Benajiba, Yacine, Rosso, Paolo et Miguel BenediRuiz, José. 2007. « ANERsys : An Arabic Named Entity Recognition System Based on Maximum Entropy ». p. 143-153.

Blachère, Régis et Gaudefroy-Demombynes Maurice. 1966. *Grammaire de l'arabe classique*. Paris : Maisonneuve et Larosse.

Boudchiche, Mohamed et Mazraoui, Azzeddine. 2016. « Approche hybride pour le développement d'un lemmatiseur pour la langue arabe ». In 13th African Conference on Research in Computer Science and Applied Mathematics, Hammamet, Tunisia, p. 147.

Cohen, David. Consulté le 12 janvier 2017. *Arabe (monde)–langue*. Encyclopedie Universalis.

Debili, Fathi, Hadhémi, Achour et Souissi, Emna. 2002. « La langue arabe et l'ordinateur : de l'étiquetage grammatical à la voyellation automatique ». Correspondances. 71. p. 10-28.

Ghoul, Dhaou. 2013. « Développement de ressources pour l'entraînement et l'utilisation de l'étiqueteur morphosyntaxique TreeTagger sur l'arabe ». TALN-RECITAL, pp. 17-21.

Hajic, Jan, Smrz, Otakar, Zemanek, Petr, Šnaidauf, Jan et Beska, Emanuel. 2004. « Prague arabic dependency treebank : Development in data and tools ». In Proc. of the NEMLAR Intern. Conf. on Arabic Language Resources and Tools, pp. 110-117.

Kammoun Nouha, Chaaben, Belguith, Lamia, Hadrich et Ben Hamadou, Abdelmajid. 2010. The MORPH2 new version : A robust morphological analyzer for arabic texts. In JADT 2010 : 10th International Conference on Statistical Analysis of Textual Data.

Pasha, Arfath, Al-Badrashiny, Mohqmed, Diab, Monia, Kholy, Ahmed, Eskander, Ramy, Habash, Nizar, Poolerry, Manoj, Rambow, Owen et M. Roth, Ryan. 2014. « MADAMIRA : A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic ». European Language Resources Association (ELRA).

Smrz, Otakar, Pajas, Petr. Zabokrtsky, Zdenek, Hajic, Jan, Mirovsky, Jiri et Nemec, Petr. 2007. « Learning to use the Prague Arabic Dependency

Treebank ». Amsterdam studies in the theory and history of linguistic science 4, 289, 77.
Smrz, Otakar, Šnaidauf, Jan et Zemanek, Petr. 2002. « Prague dependency treebank for arabic : Multi-level annotation of arabic corpus ». In Proc. of the Intern. Symposium on Processing of Arabic, p. 147-155.

Abstract

We propose a new tool for the morphosyntactic parsing of Modern Standard Arabic based on the linguistic study of schemes. Our first objective is to design linguistic rules for the morphosyntactic parsing and evaluate our method using the annotated corpus Arabic-PADT UD Treebank. This corpus has been initially developed to help the implementation of tools using machine learning and contains many occurrences tagged by the X label that indicate unknown classes. The second objective of our study is to propose a method to improve this resource by annotating these occurrences in the corpus. The results of the two evaluations show that our methods obtain F-measure of 0.92 and 0.90 for the morphosyntactic parsing.

Modeling Legal Terminology in SUMO

Jelena Mitrović*, Adam Pease**,
Michael Granitzer***

*University of Passau, Germany
jelena.mitrovic@uni-passau.de
http://jelena.mitrovic.rs/
**Articulate Software, USA
apease@articulatesoftware.com
www.adampease.org/professional
*** University of Passau, Germany
michael.granitzer@uni-passau.de
www.fim.uni-passau.de/data-science

Abstract. In this paper, we discuss ontological modeling of legal terminology in SUMO (Pease, 2001) and utilizing its close connection to the lexical-semantic network WordNet (Fellbaum, 1998) in the legal domain. Formal systems that allow for automated semantic interpretation of law supported by lexical resources can bring forth solutions to legal reasoning tasks. We wish to formalize legal issues in a computable language, using SUMO and its higher-order language to capture the semantics of the legal domain, which in itself is relying on many quasi-logical rules. SUMO is a formal ontology that had 1000 terms and 4000 axioms in the beginning, using those terms in SUO-KIF statements that describe the world. This set has since been expanded to more than 20,000 terms and 80,000 axioms. In this paper, we show how and why legal terms have been represented in SUMO. We also sketch out our future work on extending legal terminology in SUMO.

1. Introduction

Formal systems that allow for automated semantic interpretation of law sup-ported by lexical resources can provide solutions to many tasks related to legal reasoning. We wish to formalize legal issues in a computable language, using SUMO and its higher-order language to capture the semantics of the legal domain, which in itself relies on many quasi-logical rules.

The Suggested Upper Merged Ontology (SUMO) (Niles and Pease, 2001 ; Pease, 2001) is a formal ontology stated in SUO-KIF (Pease, 2003). SUMO began as an ontology of 1000 terms and 4000 axioms using those terms in SUO-KIF statements that describe the world. This set has since been expanded to more than 20,000 terms and 80,000 axioms. SUMO has been used in many academic and business applications, ranging from digital media (Pease and Rust, 2008), e-commerce, large scale theorem proving in first-order (Pease *et al*., 2010) and higher order logic and natural language processing (Pease and Li, 2010). SUMO is a general purpose, upper level ontology but also a collection of domain specific ontologies, all of which are integrated and logically consistent with each other. Although SUMO is large, the scope of human knowledge is such that every domain cannot be covered in advance.

Mapping SUMO to WordNet (Fellbaum, 1998) was performed nearly at the beginning of the SUMO project, as a first validation step. The mappings are continually updated and improved as the SUMO is expanded. This connection between the two resources remains valuable, and our goals is to utilize it, taking into account the intricacies of the legal domain.

The remainder of the paper is organized as follows : Section 2 covers Related work on legal ontologies in relation to SUMO ; we focus on Deontology and deontic statements in Section 3 ; we elaborate on Legal interpretation in Section 4 ; we show more existing legal SUMO formalizations in Section 5 and the connection to WordNet and future work is outlined in Section 6.

2. Related work

We want to capture an ontology of law. There has been work in AI and philosophy in modeling law and legal reasoning. Unfortunately, this effort has apparently been of two kinds. First, in philosophy, we have papers that describe a conceptual framework for law and legal reasoning. They may elucidate important issues, but they are not concerned with formalizing those issues in a computable language. Second, in AI and computer science we have attempts to model law, but the work suffers from two issues : **(1)** law is considered independently of the rest of the world, and yet since law touches on virtually every aspect of our world, little practical reasoning can be done in such a system because there is so much that is left outside the representation, and **(2)** that such work has typically employed logics that are incapable of representing the full semantics of a domain that relies heavily on higher-order logical issues such as belief and intent. Such work typically just elucidates a

taxonomy, which then must be used informally, with most of the semantics in the intuition of the user, rather than expressed in a formal system and available for automated computation.

Our present work addresses these issues by studying legal ontology in the context of an existing large ontology – SUMO. It also addresses computational sufficiency by being written in a higher order logic, so that we can at least have the capability of automated reasoning about beliefs, intent, temporal qualifica-tion, modality and many other issues that cannot be captured explicitly in a less expressive logic. The computational language used also has an implementation in particular automated theorem provers (Benzmüller and Pease, 2010a). This is not to say of course that much existing work about law in AI and philosophy is not relevant, simply that it is not sufficient for a computational theory of law. We aim to get closer to such sufficiency in this present work.

There are four often-cited conceptualizations of law in AI (a) McCarty's LLD (McCarty, 1989), (b) Stamper's NORMA (Stamper, 1991), (c) Valente's Functional Ontology of Law (Valente, 1995) and (d) the Frame-based Ontology of Van Kralingen and Visser (Visser *et al*., 1997). It is encouraging that there has been some work in modeling law in more expressive logics, although by using specialized logics, typically one issue at a time, like deontic force or temporal qualification is addressed at the expense of other issues that require expressivity greater than first-order logic. Although there are likely many issues that must be explored in combining specialized logics in a higher-order framework (Benzmüller and Pease, 2010b) that does not lessen the need to do so.

A key challenge of legal modeling is that the law itself already contains a semi-formal model of the world (Prakken and Sartor, 2015). It is a system of rules designed to govern human action. A formal system must take into account not only these rules, but their possible interpretation, since much of legal reasoning in the real world involves how to interpret the rules of law, rather than just applying them mechanically in a formal system. As such we must also be concerned with modeling argumentation and denotation (Mitrović *et al*., 2017).

There is a great deal of existing work in modeling of law. Domain-specific legal ontologies, pertaining to a certain legal practice or a type of law exist. Our goal, on the other hand, is to enrich SUMO with terms and related concepts that can be used in any legal situation.

The LRI-Core Legal Ontology (Breuker *et al*., 2002) is a core ontology that covers the main concepts that are common to all legal domains. It starts with four main categories: physical concepts, mental concepts, roles, and abstract concepts. The notion of 'what can happen to an object' is pursued.

LKIF-Core ontology (Hoekstra *et al*., 2007) consists of 15 modules, each of which describes a set of closely related concepts from both legal and common sense domains. It is therefore rather a library of ontologies relevant for the legal domain than a monolithic body of definitions. The most abstract concepts are defined in five closely related modules: top, place, mereology, time and space-time. LKIF's top ontology is largely based on the top-level of LRI-Core but has less ontological commitment in the sense that it imposes less restrictions on subclasses of the top categories.

In (Ajani et. al, 2016), a lightweight ontology related to linguistic patterns that denote legal concepts in several languages spoken in the European Union (EU) is described. A first draft of a legal ontology on the new EU regulation GDPR, with a goal of providing a legal knowledge modeling of the privacy agents, data types, rights and obligations was presented in (Palmirani *et al*., 2018). An ontology of criminal law was presented in (Gosh at al., 2017).

3. Deontology

A deontic statement makes a commitment to the way things *ought to be*. Law is fundamentally a deontic construct. (Visser and Bench-Capon, 1998) have four deontic categories – permitted (P), forbidden (F), obligatory (O), and enabled (E). (Hohfeld, 1913) provides a framework for law in the quotes from his paper below. He also defines four deontic relations which are rights-relations:

- *right* – "if X has a right against Y that he shall stay off the former's land, the correlative (and equivalent) is that Y is under a duty toward X to stay off the place"
- *duty* – "A duty or a legal obligation is that which one ought or ought not to do. `Duty' and `right' are correlative terms. When a right is invaded, a duty is violated." "… when it is said that a given privilege is the mere negation of a duty, what is meant, of course, is a duty having a content or tenor precisely opposite to that of the privilege in question."
- *privilege* – "whereas X has a right or claim that Y, the other man, should stay off the land, he himself has the privilege of entering on the land; or, in equivalent words, X does not have a duty to stay off." "… if

A has not contracted with B to perform certain work for the latter, A's privilege of not doing so is the very negation of a duty of doing so." A privilege is freedom from obligation. "The privilege of entering is the negation of a duty to stay off."

- *no-right*.

These correspond to SUMO's existing DeonticAttributes[1]. Hohfeld's notion of *duty* corresponds to SUMO's Obligation, *right* corresponds to Permission and *no-right* corresponds to Prohibition. The one missing concept appears to be the notion of *privilege*. To take Hohfeld's examples in SUMO terms, X has Permission to use or occupy his own land. There is a Prohibition against others trespassing on X's land. There is an Obligation for others to be off of X's land. The notion of privilege seems more difficult it seems to be a right that is temporary that others do not have. It's possible for a right not to be exclusive. Every person could have the right to occupy a public park, but it makes no sense to talk about everyone having the same privilege. If everyone has it, it's no longer a privilege.

Hohfeld also captured the notion of power-relations, which are "power", "subjugation", "disability" and "immunity". If Y makes a promise to X then X has the power to create a right by accepting the promise.

We also have to distinguish these notions from discussions of rights and privileges under natural law for which one view is that certain rights are inalienable, such as declared in the preamble to the US Constitution. Under this approach, a privilege can be revoked, but a right cannot. But of course under some circumstances rights can be limited or revoked. The right to liberty could be revoked after arrest and conviction of a crime. One theory is that natural rights are not derived from an organizational authority. People deserve liberty even if there are just two people in the world. No just person deserves to be imprisoned by another. In contrast, the right to one's property only holds if there is some authority that can designate or certify property ownership. A collective society may not even have a notion of property, so there can be no natural right to property. But then do we condition natural rights only by whether a society exists that does not have those rights? If there exists a non-free society do we then say that freedom is not a natural right?

1 Note that all terms in monospaced font should be interpreted as terms from the SUMO ontology. The reader can explore the full definitions for these terms on-line by entering the term in the "KBterm" field at http://sigma.ontologyportal.org : 8080/sigma/Browse.jsp

Let's look at each of these in more detail. If there is an Obligation to do something, then there is necessarily not the Permission not to do it. To phrase it as a positive example, if you have an Obligation to pay your taxes, you don't have Permission not to pay your taxes. More formally:

```
(<=>
    (modalAttribute ?FORMULA Obligation)
    (not
        (modalAttribute
            (not ?FORMULA) Permission)))
```

Also, if one has an Obligation to do something, one necessarily has Permission to do it as well.

```
(=>
    (modalAttribute ?FORMULA Obligation)
    (modalAttribute ?FORMULA Permission))
```

The next axiom says that everything for which there is a Prohibition does not have Permission, and vice versa. At the risk of creating confusion, we should explain a bit further if the axiom does not fit with one's intuition, especially for someone used to logic programming. One might think of the phrase referring to the Cold-war communist era law that "*Everything that isn't expressly prohibited is forbidden*", but that is not what this axiom says. The logic used in SUMO does not have the same negation-by-failure property as with logic programming. Only if it is expressly the case that there is not Permission for something then it is entailed that there must be a Prohibition against that thing and vice versa. If there is simply an absence of Permission, that's not the same as an explicit assertion of the negation of Permission.

```
(<=>
    (modalAttribute ?FORMULA Prohibition)
    (not
        (modalAttribute ?FORMULA Permission)))
```

A DeonticAttribute makes a certain commitment to a state of the world. It doesn't make sense to have or be granted Permission to do something impossible. DeonticAttributes are hereby related to the AlethicAttributes of Possibility and Necessity.

```
(=>
    (modalAttribute ?FORMULA Permission)
    (modalAttribute ?FORMULA Possibility))
```

4. Legal Interpretation

In (Prakken and Sartor, 2015) the authors note a number of terms which are subject to interpretation in law: "duty of care", "misuse of trade secrets" or "intent", and qualified with general exception categories, such as "self-defense", "force majeure" or "unreasonable". We must be able to model the law, the interpretation of the law, and the application of the law by its authorities. SUMO contains the notion of an Argument and one subclass is LegalOpinion which can have premises and conclusions which relate a Proposition to the Argument. This allows for extensions of SUMO related to Legal argumentation, and also to tasks that have to do with persuasiveness and rhetoric (Mitrović *et al*., 2017).

Because law must be interpreted in the real world, and it will rarely be possible to enumerate all possible contextual impacts in advance, as part of a statute, it is helpful to have a notion of *defeasible reasoning* and work in AI and law has often chosen such formalisms. Thus, each rule becomes a template or default that may be modified by other rules or circumstances to be enumerated later.

A key feature of interpretation under many systems of law is that of law being an adversarial system – each party to a case is charged with pushing their interpretation to be the most favorable to themselves, and a supposedly neutral party – the judge or jury – is charged with deciding between competing arguments. This may lead each party to widely different interpretations of statutes or even single words. One goal of formal legal reasoning therefore must be to allow for a range of different interpretations. Some actions may be strictly allowed or prohibited but others will fall in a range of relative likelihoods.

Any model of law must account for the roles of many agents and institutions in interpretation and application. In most Western systems, there is the notion of at least three branches of government that all have a role to play – the *legislative branch* that writes laws, the *judicial branch* that administers those laws, and the *executive branch* that enforces the laws. There is also a relationship between formal, codified law and social convention and morality. Each is a reflection of the other, to some degree. For example, having a sexual relationship with a person other than one's current spouse would likely be contrary to the moral views of many, but there may be at least some exceptions, for example while still being legally married to a spouse who is missing and has been presumed dead for years. In some countries and time periods

such actions would also be criminal and subject to legal sanction. As moral views have changed over time, so have legal frameworks, and behaviors many might still say are immoral are no longer criminal.

We must have a framework that captures context and intent. A person that plans a murder will face a tougher legal sanction than one who follows a momentary violent impulse or an action that could not be known in advance to result in death, under most systems of law. A person who drinks and drives may face a stiffer sanction for the same accident as one in which the party at fault is sober, because one should know that drinking impairs safe driving. On the other hand, a child is often found less culpable for same action as an adult, because a child has less ability to understand the implications of his or her actions.

Law can also involve a degree of consent, and consent itself can alter a legal interpretation. A child may be presumed not to understand the consequences of his actions, and so any consent given may be disregarded. By entering a particular country one is given consent to be subject to its laws.

Games and competitions also involve sets of rules, and when in the context of professional sports can also be matters of law. By entering into a competition one is either informally presumed to give consent to the rules, or may do so as a matter of formal contract. Someone caught cheating on a golf course in a match between friends would be subject to disapproval. A professional who colludes with a colleague to move a ball and change the outcome of a match might be subject to forfeit a prize in at least a civil prosecution. Throwing a match when there are bets on the outcome could result in criminal racketeering charges.

Consent and conformance also has a legal impact when it comes to implementation of legal sanction. A person may pay a parking ticket by check rather than having his bank comply with a court order to pay a fine out of his bank account without his consent. A person may report to jail rather than being carried bodily into a cell by officers. The legal system may reward the lawbreaker who is compliant with his incarceration with "time off for good behavior" rather than a lengthier sentence after attempted escape or rules-breaking within prison. The acceptance of guilt and appearance of contrition is part of the context of punishment.

5. Existing Formalization in SUMO

SUMO consists of many terms and definitions regarding Law. Some of them are the DeonticAttributes, such as Obligation, Permission and Prohibition, which correlate to the way things ought to be, and AlethicAttributes (denoting modalities of truth) such as Possibility and Necessity. In addition to the type hierarchies found in most ontologies, SUMO has expressive definitions in higher order logic that can be used for complex reasoning with a suitably expressive theorem prover, such as LEO-II (Benzmueller and Pease, 2010), for example that conferring Permission entails not conferring a Prohibition for the same thing, or that having a legal agreement with permission to do something means that in fact one has the right to do such a thing, as it can be stated in SUMO bellow:

```
(=>
    (confersNorm ?E ?F Permission)
    (not
        (confersNorm ?E ?F Prohibition)))
(=>
    (agreementClause ?PROP Permission ?AGREEMENT ?AGENT)
    (holdsRight
        (exists (?PROC)
            (and
                (realization ?PROC ?PROP)
                (agent ?PROC ?AGENT))) ?AGENT))
```

The existing SUMO model has most of the elements needed for a legal framework. It is implemented in a higher-order language, which, unlike a description logic or even first-order logic, allows us to use entire formulas as arguments to relations. This enables us to encode statements about belief, intent and modality. For example:

```
(=>
    (and
        (holdsDuring ?T
            (desires ?M
                (attribute ?V Dead)))
        (instance ?MURDER Murder)
```

(agent ?MURDER ?M)
(patient ?MURDER ?V)
(earlier ?T (WhenFn ?MURDER)))
(attribute ?MURDER Premeditated))

If the perpetrator committed a murder and wanted the victim dead before the actual murder was committed, then it was considered to have been premeditated. Rather than just employing a named term in a taxonomy, we define a rule that can be used in an automated theorem prover, without recourse to human intuition, to compute answers to a problem such as the appropriate sentence for a premeditated murder. It includes higher-order constructs – first in the use of temporal logic that qualifies a formula:

(desires ?M
(attribute ?V Dead))

as being true – that it holdsDuring - time ?T. The temporally qualified formula itself is higher order, consisting of a propositional attitude – a desire held by an agent that a particular formula should be true.

The expressiveness of the logic used allows us to capture the full semantics of what premeditated murder is, rather than just creating a taxonomy where *PremeditatedMurder* is a kind of *Murder* and each has only an English definition that cannot be employed for logical reasoning.

One area of SUMO that is in need of expansion, especially in regards to legal ontology, is that of CaseRole(s), which specify the roles that entities play in particular events and processes. The CaseRole of patient is particularly overloaded. We have specialized terms for an arrested party, targetInAttack and defendant but many more are needed. While it remains to be seen whether CaseRole(s) are needed for each of the following, we will need some way to distinguish from the more generic patient, the following: evidence, judge, witness, arrestingOfficer, jailer, defenseLawyer, victim, perpetrator, and prosecutor.

More relevant existing legal terms in SUMO, and in some cases, their hierarchical relationships are shown in Annex 1 of this paper.

6. Future Work

Ontological modeling of legal terminology in SUMO in combination with the lexico-semantic database WordNet (Fellbaum, 1998) is an area of active

research for us. WordNet is a general-purpose resource and similarly does not contain all specialized and technical terms used in English. While they have some basic legal notions, many more terms are needed in WordNet and in SUMO to cover common legal terms, especially Latin terms. For example, WordNet has a synset for "*de jure*" encoded as an adjective and also as an adverb. It is mapped to SUMO's NormativeAttribute in both cases[2].

NormativeAttribute is a SUMO Class containing all of the Attributes that are specific to morality or legality. Many of these attributes express a judgement that something ought or ought not be the case. This mapping is however not specific enough. SUMO needs a term and logical expressions to define the notion of "*de jure*" actions, which are more specific than the broad notion of any NormativeAttribute.

We will also rely on WordNet Domains (Bentivogli *et al*., 2014) to obtain the WordNet 3.1 synsets from the legal domain and their mappings to SUMO. We will map more legal domain synsets to SUMO and more fully formalize related legal terminology in SUMO.

Another line of future research has to do with the extensions of SUMO relevant for the field of Legal argumentation and Legal discourse analysis where SUMO was already used (Pease *et al*., 2019). We foresee many interesting applications, such as the analysis of persuasiveness in line with ontological models of rhetorical figures as presented in the RetFig ontology of rhetorical figures (Mladenović and Mitrović, 2013). Legal experts will be involved in evaluation of such systems.

References

Ajani, Gianmaria, Guidoa Boella, Luigia di Caro, Livioc Robaldo, Llioa Humphreys, Sabrinaa Praduroux, Piercarlob Rossi, and Andrea Violato. 2016. "The European Legal Taxonomy Syllabus: A multi-lingual, multi-level ontology framework to untangle the web of European legal terminology." Applied Ontology,11 (4): 325-375.

Bentivogli, Luisa, Pamela Forner, Bernardo Magnini, and Emanuele Pianta. 2014. "Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing." Proceedings of International Conference on Computational Linguistics COLING, Geneva, Switzerland, 101-108.

2 http://sigma.ontologyportal.org:8080/sigma/WorNet.jsp?simle=null&kb=SUMO&lang=Eng-lishLanguage&flang=SUO-KIF&word=de+jure&POS=0

Benzmüller, Christoph, and Adam Pease. 2010a. "Progress in automating higher-order ontology reasoning." In Boris Konev, Renate Schmidt, and Stephan Schulz, editors, Workshop on Practical Aspects of Automated Reasoning (PAAR-2010). CEUR Workshop Proceedings, Edinburgh, UK, July 14.

Benzmüller, Christoph, and Adam Pease. 2010b. "Reasoning with Embedded Formulas and Modalities in SUMO." The ECAI-10 Workshop on Automated Reasoning about Context and Ontology Evolution.

Breuker, Joost, Abdullatif Elhag, Emil Petkov Winkels, and Radboud Winkels. 2002. "Ontologies for legal information serving and knowledge management". In : T. Bench-Capon, A. Daskalopulu and R. Winkels (eds.), Legal Knowledge and Information Systems. Jurix 2002 : The Fifteenth Annual Conference, 73-82. Amsterdam : IOS Press.

Fellbaum, Christiane.1998. "WordNet : An Electronic Lexical Database." Cambridge : MIT Press.

Ghosh El Mirna, Hala Naja, Habib Abdulrab, and Mohamad A. Khalil. 2017. "Ontology Learning Process as a Bottom-up Strategy for Building Domain-specific Ontology from Legal Texts." In Proceedings of the 9th International Conference on Agents and Artificial Intelligence (ICAART 2017): 473-480.

Hoekstra, Rinke, Breuker, Joost Di Bello, Marcello and Alexander Boer. 2007."The LKIF Core Ontology of Basic Legal Concepts." LOAIT.

Hohfeld, Wesley Newcomb. 1913. "Some Fundamental Legal Conceptions as Applied in Judicial Reasoning." The Yale Law Journal, 23(1):16-59.

McCarty, L. Throne. 1989. "A Language for Legal Discourse I. Basic Features." In Proceedings of the 2nd International Conference on Artificial Intelligence and Law, ICAIL. New York, NY, USA, 180-189.

Mitrović, Jelena, Cliff, O'Reilly, Miljana Mladenović, and Siegfried Handschuh. 2017. "Ontological Representations of Rhetorical Figures for Argumentation Mining." Argument & Computation, 8 (3): 267-287.

Mladenović, Miljana, and Jelena Mitrović. 2013. "Ontology of Rhetorical Figures for Serbian.", Lecture Notes in Computer Science and Artificial Intelligence 8082/Springer-Verlag Berlin Heidelberg, 386-393.

Niles, Ian, and Adam Pease. 2001. "Toward a Standard Upper Ontology." In Welty, C. and Smith, B., editors, Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001), 2-9.

Niles, Ian, and Adam Pease. 2003. "Linking Lexicons and Ontologies : Mapping WordNet to the Suggested Upper Merged Ontology." Proceedings

of the IEEE International Conference on Information and Knowledge Engineering, 412-416.

Palmirani, Monica, Michele, Martoni, Arianna, Rossi, Cesare, Bartolini, and Livio Robaldo. 2018. "PrOnto : Privacy Ontology for Legal Reasoning." In Electronic Government and the Information Systems Perspective. Springer International Publishing, 139-152.

Pease, Adam, and Godfrey Rust. 2008. "Formal Ontology for Media Rights Transactions, in Semantic Web Methodologies for E-Business Applications." Ed. Roberto Garcia. IGI publishing.

Pease, Adam, and John Li. 2010. "Controlled English to Logic Translation. In Theory and Applications of Ontology". Ed. Roberto Poli, Michael Healy, and Achilles Kameas, Springer, ISBN : 978-90-481-8846-8.

Pease, Adam, Jennifer Cheung Pease, and Andrew K.F. Cheung. 2019." Formal ontology for discourse analysis of a corpus of court interpreting." Babel, 64 (4): 594-618.

Pease, Adam, Geoff, Sutcliffe, Nick, Siegel, and Steven Trac. 2010. "Large Theory Reasoning with SUMO at CASC." AI Communications, 23(2-3): 137-144, Special issue on Practical Aspects of Automated Reasoning, IOS Press, ISSN 0921-7126.

Pease, Adam. 2011. "Ontology : A Practical Guide." Angwin, CA : Articulate Software Press. ISBN 978-1-889455-10-5.

Prakken, Henry, and Giovanni Sartor. 2015. "Law and Logic." Artificial Intelligence, 227, C : 214-245.

Stamper, Ronald K. 1991. "The role of semantics in legal expert systems and legal reasoning." Ratio Juris, 4 (2): 219-244.

Valente, Andre. 1995. "Legal knowledge engineering : A modelling approach." PhD thesis, University of Amsterdam.

Visser, Pepijn R.S., and Trevor J.M. Bench-Capon. 1998. "A comparison of four ontologies for the design of legal knowledge systems." Artificial Intelligence and Law, 6 (1) :27-57.

Visser, Pepijn R.S., Robert Willem van Kralingen, and Trevor J.M Bench-Capon. 1997. "A Method for the Development of Legal Knowledge Systems." In Proceedings of the Sixth International Conference on Artificial Intelligence and Law (ICAIL'97), Melbourne, Australia.

Résumé

Dans cet article, nous discutons de la modélisation ontologique de la terminologie juridique dans les ontologies formelles, comme SUMO (Pease, 2001) et de la possibilité d'utiliser son lien étroit avec le réseau lexico-sémantique WordNet (Fellbaum, 1998) dans le domaine juridique. Les systèmes formels qui permettent une interprétation sémantique automatisée du droit à l'aide de ressources lexicales peuvent apporter des solutions aux tâches de raisonnement juridique. Nous souhaitons formaliser les questions juridiques dans un langage informatique, en utilisant SUMO et son langage courant d'ordre supérieur pour saisir la sémantique du domaine juridique, qui en soi repose sur de nombreuses règles quasi logiques. SUMO est une ontologie formelle qui avait 1 000 termes et 4 000 axiomes au début, utilisant ces termes dans les énoncés SUO-KIF qui décrivent le monde. Cet ensemble a depuis été élargi à plus de 20 000 termes et 80 000 axiomes. Dans cet article, nous montrons comment et pourquoi les termes juridiques ont été représentés dans SUMO. Nous esquissons également les grandes lignes de nos futurs travaux sur l'extension de la terminologie juridique dans SUMO.

Annex 1

JudicialProcess **LegalAppeal** **LegalSummons** **LegalDecision** **LegalAquittal** **LegalAward** **LegalConviction** **LegalDismissal** **Sentencing** **LegalOpinion** **LegalServices**	**Law** **Legal** **LegalAction** **LegalCharge** **PlacingUnderArrest** **LegalAgent** **Corporation** **Government** **HumanAdult**
LegalSystemAttribute **AcceptsICJJurisdiction** **CivilLaw** **EnglishCommonLaw** **IslamicLaw** **JudicialReviewOfExecutiveActs** **JudicialReviewOfLegislativeActs** **NapoleonicCode** **RomanCanonLaw**	**JudiciaryFn** **judicialBranch** **SupremeCourtFn** **localSystemType** **JudicialOrganization** **RegionalLawFn** **CourtRoom** **FieldOfLaw**

Articles
session « Terminology and text mining »

Extractions de graphies terminologiques à partir de patrons morphosyntaxiques : propositions et comparaisons

Amaury Delamaire*, Michel Beigbeder**,
Mihaela Juganaru-Mathieu***

Mines Saint-Étienne, Univ Lyon,
Univ Jean Monnet, IOGS, CNRS,
UMR 5516 LHC, Institut Henri Fayol
F - 42023 Saint-Étienne France
*amaury.delamaire@emse.fr
**mbeig@emse.fr
***mathieu@emse.fr

Résumé. Les patrons morphosyntaxiques décrivent des séquences d'étiquettes éponymes. Ils sont particulièrement utilisés pour la construction de terminologies. Quelques patrons sont exploités de manière récurrente dans la littérature. Nous proposons ici une extension d'un de ces patrons ainsi qu'un protocole d'évaluation. Les expériences effectuées sur la base de ce protocole mettent en évidence l'utilité de ce nouveau patron. L'intérêt du contexte dans les calculs de métriques terminologiques (ici la NC-valeur) est également investigué.

1. Introduction

L'augmentation permanente de la production de données textuelles implique une nécessité de les organiser. Le traitement automatique du langage (TAL) propose des solutions relatives à cette problématique qui est très vaste. Nous nous intéresserons plus particulièrement à l'extraction automatique de terminologie à partir des documents techniques et scientifiques.

1.1. Le Traitement Automatique du Langage

L'appellation TAL regroupe l'ensemble des processus d'analyses automatiques portant sur la langue, tant parlée qu'écrite. La complexité des différents processus d'analyses diffèrent fortement selon la nature de la tâche. Cette

complexité émane notamment du besoin d'experts pour la construction de ressources et/ou l'évaluation de résultats. L'automatisation de la construction de terminologies est particulièrement touchée par cette limitation, comme nous le verrons en partie 1.4. La complexité provient également du faible consensus concernant la terminologie : alors que tout le monde s'accorde à annoter "réseau" comme un nom commun, la même chose ne peut être dite concernant sa qualification en tant que terme.

Les tâches complexes du TAL sont également impactées par la propagation d'erreurs issues de sous-tâches moins complexes. Des sous-tâches comme l'étiquetage morphosyntaxique ou la lemmatisation produisent peu d'erreurs, mais les erreurs produites se propagent aux tâches plus complexes comme l'extraction terminographique qui en retour produisent davantage d'erreurs.

1.2. L'analyse terminologique

Une terminologie désigne une ressource linguistique organisée listant des termes. Les termes correspondent à des concepts dont le sens est spécifique à une thématique particulière.

> Exemple : *Réseau de neurones* est terminologique en médecine et en apprentissage automatique, mais avec des sens distincts.

Dans le cadre d'une construction manuelle, la distinction entre termes et non termes est faite par un ou plusieurs experts. Dans le cadre d'une construction automatique, la distinction est produite par un algorithme qui tente d'estimer le potentiel terminologique d'un terme candidat donné. Des approches hybrides ont été proposées, comme nous le verrons en partie 2.

La subjectivité demeure une problématique récurrente : la distinction terme/non terme peut varier fortement d'un expert à l'autre, selon sa perception de l'aspect terminologique. Cela explique le peu de ressources terminologiques à disposition à des fins de construction des outils d'automatisation.

Il n'en demeure pas moins qu'automatiser le processus depuis l'extraction jusqu'à l'évaluation permettrait de comparer aisément différentes méthodes. Nous le verrons en partie 2, la littérature propose différentes descriptions des "termes acceptables", descriptions figées induisant nécessairement du silence et/ou du bruit que nous analyserons par la suite.

Indépendamment du niveau d'automatisation de l'analyse, la construction de terminologie passe nécessairement par une étape d'extraction terminographique.

1.3. Automatisation de l'évaluation terminologique

L'extraction terminographique consiste en l'extraction de graphies aboutissant à la construction de termes candidats. Un terme candidat peut se définir comme un élément potentiel de la terminologie, pouvant regrouper plusieurs graphies issues de l'extraction terminographique. Plusieurs car celles-ci ne sont pas nécessairement canoniques, i.e. elles peuvent être des variantes graphiques ou sémantiques d'un même terme. Dans le corpus ACL-RD-TEC 1.0[1], les graphies sont annotées, mais les candidats termes ne sont pas construits : les termes apparaissent sous diverses formes fléchies.

> Exemple : *neural network* vs. *neural networks, train algorithm* vs. *trained algorithm,* etc.

L'extraction terminographique est l'étape préliminaire à la construction de termes candidats. Comme nous le verrons en partie 2.1, les patrons morphosyntaxiques sont régulièrement utilisés pour automatiser le processus bien qu'il n'y ait cependant pas encore de prééminence particulière d'un patron sur les autres. Les patrons permettent d'identifier les graphies qui correspondront à des termes candidats, les termes candidats sont ensuite filtrés selon différentes méthodes, afin de ne conserver que les "vrais termes". Le filtrage est généralement effectué par un expert qui valide manuellement la liste triée extraite automatiquement par un système. L'extraction produisant généralement de très nombreux termes candidats, le tri de la liste selon diverses métriques a vocation à faciliter le travail de l'expert. De même que divers patrons sont proposés, diverses métriques de potentiel terminologique ont été introduites dans la littérature.

L'automatisation de la construction de terminologies pose cependant la problématique de l'évaluation et de la comparaison des systèmes.

1.4. Les évaluations de systèmes automatisés

La construction de terminologies, automatique ou non, pose le problème de son évaluation. Comme évoqué en partie 1.2, la construction de terminologies est impactée par la perception des experts qui participent à sa construction. La problématique est autre dans un contexte d'automatisation du processus : les experts ne participent plus à la construction (sauf exceptions, cf. partie 2.1) mais à l'évaluation. Des solutions ont été proposées, comme nous

1 http://pars.ie/lr/acl-rd-tec-terminology

le verrons en partie 2.2, mais la problématique reste ouverte. L'apparition de corpora annotés au niveau terminologique, bien que rares, va permettre de comparer des systèmes de manière directe et pertinente (cf. partie 3.3) modulo certains biais introduits par lesdites ressources (cf. partie 3.4). Les résultats présentés ici suivent cette méthode d'évaluation.

Nous commencerons par présenter diverses techniques exploitées à ces fins en partie 2. Nous présentons ensuite les modifications proposées ainsi que les justifications qui les appuient en partie 3. Suivent les résultats d'une étude comparative entre plusieurs méthodes dans la même partie.

2. État de l'art

Nous nous focalisons ici sur l'extraction terminographique, spécialement centrée sur l'extraction à partir de patrons morphosyntaxiques (partie 2.1). Suit une présentation des différentes méthodes d'évaluation des extractions proposées dans la littérature (partie 2.2). Enfin, nous introduisons le contexte de nos expériences relativement à ces constats (partie 2.3).

2.1. Extractions à partir de patrons morphosyntaxiques

De nombreuses expériences sur l'automatisation de l'extraction terminographique exploitent les patrons morphosyntaxiques. Ces derniers peuvent se définir comme des séquences d'étiquettes morphosyntaxiques correspondant à des chaînes de caractères à fort potentiel terminologique. Le premier patron a été introduit par Justeson et Katz (Justeson et Katz, 1995) :

(Patron 1) :

((JJ[RS] ?**|NNS** ?**)**+**((JJ[RS]** ?**|NNS** ?**)*****(NNS** ? **IN)**?**)(JJ[RS]** ?**|NNS** ?**)*****)NNS** ?

Où : **JJ** est un adjectif, potentiellement comparatif (**JJR**) ou superlatif (**JJS**), **NN** un nom commun potentiellement au pluriel (**NNS**), **IN** une préposition.

Le jeu d'étiquettes utilisé pour le patron 1 et les autres est celui du Penn TreeBank[2]. Bien que le patron proposé a été largement réduit/complété dans diverses expériences, les propriétés introduites par Justeson et Katz restent valables :

2 https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

1. la majorité des chaînes de caractères correspondant à un terme sont des séquences de noms communs (>90 %),
2. l'insertion d'un mot supplémentaire entre les mots constituants un terme est difficile/impossible,
3. la majorité des termes fait moins de cinq mots, avec un pourcentage maximal à deux mots (exception faite de certains domaines comme la médecine).

Des expériences ont été menées sur des séquences d'étiquettes autres que nominales – en l'occurrence verbales – (L'Homme, 2012) mais la complexité de la tâche reste un obstacle. Alors qu'un groupe nominal est autonome, un verbe induit une arité et la spécification de la nature de ses acteurs.

Exemple : *Mon professeur entraîne un réseau de neurones.*

Le groupe « réseau de neurones » est autonome, contrairement au verbe « entraîner » qui nécessite des spécifications supplémentaires :

- arité deux : transitif,
- nature du sujet : agent,
- nature de l'objet : algorithme d'apprentissage.

Cette complexité déjà rencontrée dans d'autres domaines du TAL (Baker *et al*., 1998) nous pousse à nous concentrer sur les groupes nominaux, plus simples à analyser, pour lesquels des recherches restent à effectuer. De fait, de nombreux autres chercheurs se concentrent sur les groupes nominaux et introduisent des modifications au patron introduit par Justeson et Katz. Ainsi, (Park *et al*., 2002) introduit la possibilité d'avoir une conjonction de coordination dans un terme candidat, de même que des formes verbales (participes passés et gérondifs) et des noms propres. (Frantzi *et al*., 2000) font quant à eux une analyse comparative entre trois patrons morphosyntaxiques, dont le patron 1 de (Justeson et Katz, 1995) et les deux suivants :

(Patron 2) : **(JJ[RS]?|NNS?){0,4}(NNS?)**

(Patron 3) : **NNS?{0,4}NNS?**

Ils aboutissent à la prééminence du patron 2, basée sur une méthode d'évaluation que nous présenterons dans la partie suivante.

Au-delà du choix du patron morphosyntaxique, les recherches se distinguent également par leur niveau d'automatisation. Comme évoqué en par-

tie 1, la perception de l'aspect terminologique peut être très subjective. De ce fait, des systèmes semi-automatiques ont été développés ; ils interagissent avec un expert qui corrige la terminologie construite permettant au système de s'adapter et de produire un résultat plus proche de celui attendu. Dans un contexte d'automatisation complète, l'aspect terminologique doit être estimé par un algorithme, et doit donc être formalisé. Nous présentons dans la partie suivante la méthode de calcul retenue.

2.2. Estimation du potentiel terminologique

L'estimation du potentiel terminologique d'un terme candidat permet de trier la liste produite par l'extraction terminographique. Nous avons retenu la méthode de la C-valeur / NC-valeur (Frantzi *et al*., 2000). Notre choix s'est porté sur cette métrique pour sa concordance avec les propriétés données par Justeson et Katz. De plus, des observations préliminaires (Delamaire *et al*., 2019) sur le corpus non annoté iSearch ont montré des résultats prometteurs. D'autres métriques sont disponibles, notamment celles basées sur une analyse contrastive entre corpus spécialisé et corpus de langue commune (Drouin, 2003 ; Peñas *et al*., 2001, Navigli et Velardi, 2002 ; etc.). Les problématiques liées à la constitution d'un corpus de langue commune nous ont cependant menés à retenir la NC-valeur pour nos expériences.

La NC-valeur peut se décomposer en deux poids distincts : la C-valeur, et ce que nous appellerons le facteur contexte (FC). Pour les deux valeurs nous devons traiter l'ensemble des termes candidats et l'ensemble des documents qui a permis l'extraction terminographique. La NC-valeur se définit comme la combinaison linéaire de ces deux poids, pour un terme candidat donné CT nous avons :

$$NC(CT) = 0.8 \times C(CT) + 0.2 \times FC(CT)$$

$$C(CT) = \begin{cases} \log(|CT|) \times f(CT) & \text{si } |ST(CT)| = 0 \\ \log(|CT|) \times \left(f(CT) - \frac{1}{|ST(CT)|} \times \sum_{b \in ST(CT)} f(b) \right) & \text{sinon} \end{cases}$$

où: $ST(CT)$ = ensemble des termes candidats contenant CT
$|CT|$ = nombre de mots de CT
$f(CT)$ = nombre d'occurrences de CT

$$FC(CT) = \sum_{b \in MC(CT)} f(CT, b)) \times poids(b)$$

où: $f(CT, b)$ = fréquence du mot b aux côtés de CT
$MC(CT)$ = ensemble des mots des contextes d'occurrences de CT
$poids(b) = \frac{|\text{termes distincts avec } b \text{ dans leur contexte}|}{|\text{termes distincts}|}$

La C-valeur peut se décrire comme une estimation de la lexicalisation d'une graphie en se basant sur sa fréquence absolue dans le corpus. Elle prend en compte le fait qu'une graphie est incluse dans une autre graphie. Cet aspect permet aux graphies plus longues de ne pas être écrasées par les plus courtes, généralement plus fréquentes. L'hypothèse sous-jacente est que si une graphie est incluse dans d'autres graphies, cette graphie a de fortes probabilités d'être terminologique. La taille de la graphie en nombre de mots est également considérée dans le calcul de la C-valeur pour pallier en partie aux disproportions de fréquences. Plus la C-valeur est grande plus la graphie a de chance d'être terminologique.

Le facteur contexte s'appuie quant à lui sur le contexte d'apparition d'une graphie, contexte défini par les noms, verbes et adjectifs qui cooccurrent avec elle dans le texte. Un poids est attribué à chaque mot du contexte relativement au nombre de différentes graphies avec lesquelles il cooccurre.

2.3. Évaluation de terminologies

La méthode d'évaluation d'une terminologie est une problématique ouverte. La faible quantité de ressources dédiées à disposition constitue un frein important à l'automatisation de l'évaluation. Bien que de nombreuses recherches s'affranchissent d'une évaluation, des solutions ont été proposées,

notamment via l'annotation d'experts. À partir de ces annotations, (Frantzi *et al*, 1999) estiment les ratios des véritables termes en tête d'une liste triée de termes candidats relativement à la queue de la liste. Ces mesures donnent une estimation de la capacité du système à mettre en avant de « vrais » termes relativement aux non-termes.

Bien que permettant une comparaison objective entre différents systèmes, cette méthode d'évaluation ne permet pas de tirer de conclusions plus précises que des tendances en début et fin de liste. L'exploitation de ressources dédiées va nous permettre de développer cette méthode, comme nous le verrons en partie 3.3.

L'exploitation de terminologies préconstruites pose cependant le même problème que l'évaluation par des experts, à savoir la subjectivité de ses auteurs et leur propension à accepter certaines séquences morphosyntaxiques plutôt que d'autres ; nous développerons cet aspect en partie 3.4.

Nous avons présenté la méthode d'extraction terminographique qui nous concerne ici ainsi que les problématiques liées à l'automatisation de la construction terminologique. Nous avons pu constater une absence de consensus concernant l'évaluation de terminologies de même que sur le choix du patron morphosyntaxique à exploiter. Nous présentons en partie 3 nos expériences relatives à ces deux problématiques.

3. Extension de patron morphosyntaxique et évaluation

Nous présentons ici notre expérience portant sur une extraction terminographique à partir de patrons morphosyntaxiques. Nous l'avons vu en partie 2, différents patrons ont été proposés dans la littérature. (Frantzi *et al*., 2000) ont proposé une évaluation de ces patrons à partir d'annotations d'experts, permettant de les comparer. Nous proposons de compléter cette méthode d'évaluation à partir d'un corpus de référence et l'application de mesures standards utilisées dans le domaine de la recherche d'informations. Nous présentons dans la partie 3.1 le corpus exploité ainsi que les prétraitements. Dans la partie 3.2 nous complétons le « meilleur » patron identifié par (Frantzi *et al*., 2000), avant d'introduire notre méthode d'évaluation en partie 3.3. Nous concluons par une présentation des résultats de notre étude comparative en partie 3.4.

3.1. Corpora exploités

Dans un premier temps nous avons mené des expériences (Delamaire *et al*., 2019) sur le corpus en anglais iSearch, catégorisé par thème et constitué de documents techniques. En l'absence d'experts capables d'extraire manuellement ou de valider des candidats termes, iSearch ne nous a pas permis de comparer différents systèmes. Il nous a cependant permis d'observer certains silences sur des termes qui sont clairement terminologiques auxquels nous tentons de répondre dans ces expériences et que nous développons dans la partie suivante.

Nous avons sélectionné le corpus ACL-RD-TEC 1.0 comme corpus de référence pour notre expérience. Il est composé d'environ 11 000 articles scientifiques en TAL et domaines connexes où chaque article est associé à sa liste de termes. Les propriétés du corpus sont les suivantes :

- 200 méga octets de corps de textes sans balises XML,
- 34 millions de mots,
- 25 000 termes distincts annotés manuellement,
- 21 500 termes distincts composés de plusieurs mots,
- 85,8 termes en moyenne par document,
- 40,5 termes composés de plusieurs mots en moyenne par document.

Ces annotations nous permettront d'appliquer les méthodes d'évaluations standards de recherche d'information que nous présenterons en partie 3.3.

3.2. Choix et extension d'un patron morphosyntaxique

Comme évoqué en partie 2, plusieurs patrons morphosyntaxiques ont été proposés dans la littérature. Alors que certains nous semblent trop permissifs, notamment au travers des groupes prépositionnels et des conjonctions de coordination, nous proposons de compléter le patron identifié comme étant le plus pertinent par (Frantzi *et al*., 2000). En effet, ils obtiennent les meilleurs résultats avec des séquences de noms communs et adjectifs terminant par un nom. Nous proposons de compléter cette séquence avec d'autres éléments, repris notamment par d'autres chercheurs.

Nous proposons l'introduction de deux formes verbales dans le patron : les gérondifs et les participes passés. Cette modification s'explique par une similarité de comportement entre gérondifs, participes et adjectifs : ils agissent comme modifieurs d'un nom commun, le tout formant potentiellement un

terme. De plus, les étiqueteurs morphosyntaxiques confondent parfois les trois. Ajouter ces deux formes devrait théoriquement permettre de repérer des graphies pertinentes ignorées au préalable. Cela permet par exemple l'extraction de *trained model, training algorithm,* etc.

(Patron 4) : **(JJ[RS]?|NNS?|VB[GD]|NNPS?){0,4}NNS?**

Où **VB** est un verbe soit au gérondif (**VBG**) soit au participe passé (**VBD**) et **NNP** un nom propre potentiellement au pluriel (**NNPS**).

Nos expériences préliminaires sur le corpus iSearch (Lykke *et al*., 2010) nous ont rapidement mené à un constat : la présence de noms propres dans les termes, ignorés dans la plupart des patrons de la littérature. Nous avons donc introduit la possibilité d'un nom propre comme modifieur du nom commun final. L'introduction des noms propres permet de repérer des graphies telles que *Markov chain, Lie algebra,* etc. qui correspondent bien à des termes spécifiques aux mathématiques.

3.3. Méthode d'évaluation

Nous proposons une méthode d'évaluation inspirée de la recherche d'informations. La structure du corpus ACL-RD-TEC nous permet d'appliquer les métriques d'évaluation classiques en recherche d'informations. Nous l'avons vu en partie 2, une extraction terminographique automatique est généralement évaluée par sa propension à *faire remonter* des éléments effectivement terminologiques dans une liste triée selon une certaine métrique. Cette propension est généralement estimée au travers de ratios de termes effectifs calculés sur les x premiers et y derniers éléments de la liste en question, x et y étant déduits d'un seuil de fréquence. Cette restriction à $x+y$ éléments s'explique par l'évaluation manuelle, contrainte par le besoin d'experts et de temps. Cette probabilité, bien que permettant de comparer les systèmes, ne permet pas d'analyser en intégralité la distribution des *vrais* termes dans la liste triée proposée. Pour cette raison, et grâce au corpus ACL-RD-TEC, nous avons pu appliquer les métriques de recherche d'informations proposées dans les conférences TREC, pionnières dans la systématisation de l'évaluation des systèmes de recherche d'informations (Manning *et al*., 2009). Celles-ci nous ont permis de compléter les informations préliminaires fournies par l'analyse de $x+y$ éléments en tête et queue de liste triée.

Parmi les métriques proposées dans le logiciel TREC_EVAL[3], nous exploitons particulièrement les mesures de précision relatives au rappel. Ces dernières nous permettent de construire des représentations visuelles explicites de la distribution des *vrais* termes dans la liste construite – relativement au corpus ACL-RD-TEC. La F1-mesure, théoriquement pertinente, ne peut être exploitée ici : l'absence de besoin d'experts nous permet de ne pas réduire la taille de la liste de sortie, qui est de fait de taille bien supérieure au nombre de termes effectifs à extraire. En conséquence, la précision de même que la F1-mesure sont extrêmement faibles. Les mesures de précision relatives au rappel sont cependant particulièrement explicites, comme nous allons le voir dans la partie suivante.

3.4. Résultats

Nous avons comparé la qualité d'extraction de quatre patrons morphosyntaxiques associés à des calculs de C et NC-valeurs. Nous l'avons vu précédemment, la NC-valeur est une combinaison linéaire de deux poids, la C-valeur et le facteur contexte (FC). Pour chaque patron, nous avons effectué une extraction sur le corpus ACL-RD-TEC à partir d'annotations morphosyntaxiques issues de Stanford CoreNLP[4] (Manning *et al.*, 2014). Plusieurs listes triées ont ensuite été construites afin de déterminer l'incidence de l'introduction du facteur contexte.

Les cinq listes par patron ont été construites à partir des combinaisons linéaires suivantes :

(1) $0 \times$ C-valeur $+ 1 \times FC$
(2) $0.2 \times$ C-valeur $+ 0.8 \times FC$
(3) $0.5 \times$ C-valeur $+ 0.5 \times FC$
(4) $0.8 \times$ C-valeur $+ 0.2 \times FC$
(5) $1 \times$ C-valeur $+ 0 \times FC$

Le processus d'extraction produit donc cinq listes triées par patron morphosyntaxique ; chaque tri est effectué selon une combinaison linéaire de C-valeurs et FC. Pour les combinaisons 1 et 5, seul l'un des deux poids est pris en compte – respectivement la C-valeur et le FC. Une comparaison entre les patrons présentés ici met en évidence une limite de taille intrinsèque pour tous excepté le patron 1. La complexité de sa structure nous a poussés à appli-

3 https://trec.nist.gov/trec_eval/
4 https://stanfordnlp.github.io/CoreNLP/

quer la limite après l'extraction, nous permettant d'obtenir des graphies de tailles similaires pour les quatre patrons.

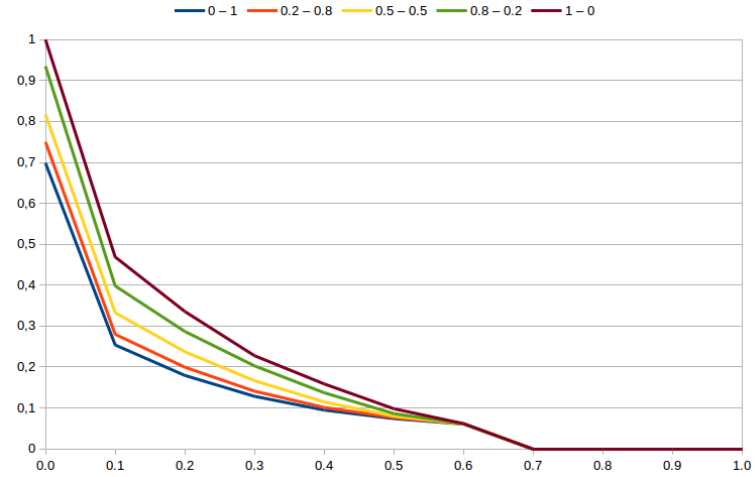


Figure 1 : Précisions du patron 4 pour diverses combinaisons de C-valeur et FC. 0-1, 0.2-0.8, 0.5-0.5, 0.8-0.2 et 1-0 sont à comprendre tels que <Coefficient C-valeur> - < Coefficient FC>

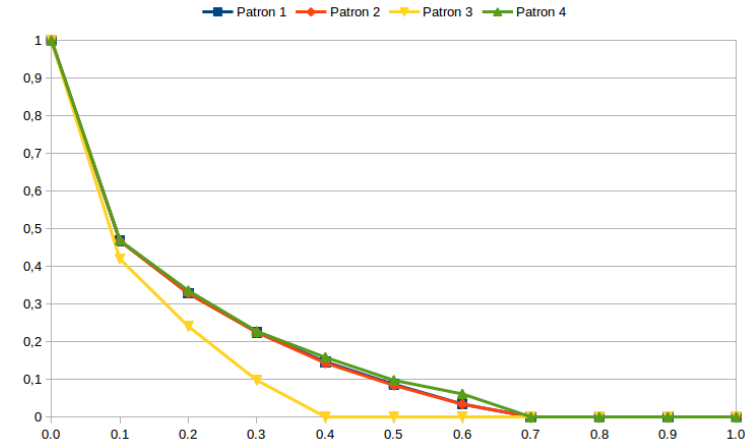


Figure 2 : Comparaison des précisions des 4 patrons.

Les figures 1 et 2 permettent d'analyser la distribution des termes et non-termes dans la liste triée construite, distinction faite sur la base des annotations du corpus ACL-RD-TEC. Avec la figure 1, nous illustrons l'impact négatif du FC avec le patron 4, les tendances sont les mêmes pour tous les autres. La figure 2 permet de comparer les quatre patrons pour la combinaison unique et optimale de C-valeur et Facteur Contexte (1 x C-valeur + 0 x *FC*). Ces courbes nous permettent d'observer plusieurs phénomènes :

- L'introduction du facteur contexte nuit systématiquement à la qualité des résultats,
- La C-valeur produit de très bons résultats relatifs au corpus ACL-RD-TEC : les premiers éléments des listes triées sont essentiellement des *vrais* termes, les derniers des non-termes,
- Trois des quatre patrons comparés ont des résultats très proches (patrons 1, 2, et 4), le quatrième est très en deçà (patron 3).
- Une légère prééminence du patron proposé (patron 4) est observable.

La faible précision de l'extraction à partir du patron 3 s'explique par l'absence d'adjectifs : alors que les termes annotés de ACL-RD-TEC en contiennent fréquemment, leur absence dans le patron provoque une « remontée » de non-termes et donc une baisse de la précision. La comparaison des quatre courbes nous permet également d'observer que les qualités d'extraction sont équivalentes pour les dix premiers pourcents de la liste construite. Cela semble indiquer que de nombreux termes annotés sont des séquences strictement nominales (**NN{2,}**), reconnues par les quatre patrons.

Le patron 1, qui inclut les groupes prépositionnels, produit des résultats proches des patrons 2 et 4 malgré le fait qu'aucun terme annoté d'ACL-RD-TEC n'en contienne. Cette proximité de performance malgré une augmentation conséquente du bruit se justifie par la méthode de tri employée. La C-valeur s'appuie sur la notion de sous-chaîne pour attribuer les scores de potentiels de terminologiques. De fait, plus les séquences sont complexes moins elles ont de chance d'être complétées dans d'autres séquences, entraînant une « remontée » des séquences plus simples aux détriments des autres. Les groupes prépositionnels n'apparaissent donc pas parmi les premiers éléments de la courbe mais dans les derniers.

A contrario, le patron 4 introduit ici permet l'extraction de nouvelles graphies pertinentes à partir de séquences morphosyntaxiques qui lui sont propres.

Exemples :

modified/VBD *interkeyboard*/JJ *dialogues*/NNS
automated/VBD *information*/NN
named/VBD *entity*/NN *matcher*/NN
binarized/VBD *PCFGs*/NNS
developing/VBG *chart*/NN
scoping/VBG *pattern*/NN

La comparaison des résultats des différentes combinaisons linéaires semble contre-intuitive. La théorie de Harris veut que le sens d'un mot est déterminable par les mots qui l'entourent, théorie corroborée par les résultats de Mikolov (Mikolov *et al*., 2013) et d'autres. A contrario, nous pouvons observer ici une dégradation constante de la qualité du tri relative à l'augmentation du facteur du FC. La raison peut être la taille de la fenêtre retenue lors du calcul des poids des mots contexte dans la FC : nous avons retenu la phrase comme fenêtre de contexte, la réduire aux mots strictement mitoyens (précédent et suivant) pourrait améliorer les résultats observés ici.

4. Conclusions et perspectives

Les patrons morphosyntaxiques sont largement utilisés pour la construction automatique de terminologie. Nous avons réalisé une étude comparative de plusieurs de ces patrons, plus un que nous proposons. Nous introduisons également une méthode d'évaluation de terminologies basée sur celles de la recherche d'informations.

Nos résultats mettent en évidence les biais introduits par l'annotation manuelle ainsi qu'une légère prééminence du nouveau patron proposé. Nous avons également pu observer une dégradation de la qualité des résultats lors de l'intégration du facteur contexte dans le calcul de la NC-valeur.

Nous avons montré qu'une utilisation des mesures de RI classiques pouvait permettre de comparer différents systèmes d'extraction de candidats termes, ce à partir d'un corpus annoté.

Cette expérience est à reproduire sur d'autres terminologies préconstruites et avec davantage de patrons. Les fenêtres de calculs des facteurs contextes dans la NC-valeur sont également à varier afin d'en analyser l'impact.

Références

Baker, Collin F., Charles J. Fillmore, and John B. Lowe. 1998. « The Berkeley Framenet Project. » *Proceedings of the 17th international conference on Computational linguistics-Volume 1*. Association for Computational Linguistics.

Delamaire, Amaury, Michel Beigbeder, and Mihaela Juganaru-Mathieu. 2019. « Exploitation de syntagmes dans la découverte de thèmes. » *CORIA : Conférence en Recherche d'Information et Application*.

Drouin, P. (2003). « Term extraction using non-technical corpora as a point of leverage ». Terminology, 9(1) : pages 99-115.

Frantzi, Katerina, Sophia Ananiadou, and Hideki Mima. 2000. « Automatic recognition of multi-word terms : the C-value/NC-value method. » *International Journal on Digital Libraries* 3.2 : 115-130.

Justeson, John S., and Slava M. Katz. 1995. « Technical terminology : some linguistic properties and an algorithm for identification in text ». *Natural Language Engineering* 1.1 : pages 9-27.

Lykke, Marianne, Larsen, Birger, Lund, Haakon, and Ingwersen, Peter. 2010. « Developing a test collection for the evaluation of integrated search ». In *European Conference on Information Retrieval*, pages 627-630. Springer.

L'Homme, Marie-Claude. 2012. « Le verbe terminologique : un portrait de travaux récents. » *SHS Web of Conferences*. Vol. 1. EDP Sciences.

Manning, Christopher D., Raghavan, Prabhakar, & Schütze, Hinrich 2009. « Chapter 8 : Evaluation in information retrieval ». *Introduction to information retrieval*, pages 139-161. Cambridge University Press.

Manning, Christopher D., Surdeanu, Mihai, Bauer, John, Finkel, Jenny, Bethard, Steven J., and McClosky, David. 2014. « The Stanford CoreNLP Natural Language Processing Toolkit ». In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, pages 55-60.

Mikolov, Tomas, *et al*. 2013. « Distributed representations of words and phrases and their compositionality. » *Advances in Neural information Processing Systems*. 2013.

Navigli, R. and Velardi, P. 2002. « Semantic interpretation of terminological strings ». In Proc. 6th Int'l Conf. Terminology and Knowledge Eng, pages 95-100.

Park, Youngja, Roy J. Byrd, and Branimir K. Boguraev. 2002. « Automatic glossary extraction : beyond terminology identification. » *COLING 2002 : The 19th International Conference on Computational Linguistics*.

Peñas, A., Verdejo, F., Gonzalo, J., *et al*. 2001. « Corpus based terminology extraction applied to information access ». In *Proceedings of Corpus Linguistics*, volume 2001, page 458-465.

Abstract

Morphosyntactic patterns describe sequences of morphosyntactic tags. They are frequently used for terminology construction. Few patterns are used throughout specialized literature. We propose here an extension to one of these patterns as well as an evaluation protocol. Experiments based on this evaluation demonstrate the relevance of the pattern we introduced. Influence of context information in termhood computation (NC-value here) is also investigated.

Chinese Word Segmentation with External Lexicons on Patent Claims

Yixuan Li*, Kim Gerdes**

*LPP (CNRS), Université Sorbonne Nouvelle - Paris 3,
19 Rue des Bernardins, 75005 Paris,
yixuan.li@sorbonne-nouvelle.fr
**Almanach (INRIA), 2 Rue Simone IFF, 75012 Paris,
kim@gerdes.fr

Abstract. This paper aims to compare the performance of different Chinese word segmenters in specialized technological domains as well as to evaluate the contribution of an external lexicon to their improvement. As we are interested in patent texts whose automatic analysis is of economic and scientific importance, we attempt to tackle the hardest source text for terminology extraction in terms of language and genre : Chinese patent claims. Some previous work on Chinese segmentation adaptation to patents are based on training a new model or using predefined term dictionaries, and none focuses on the adaptability of existing state-of-art segmenters. Our approach uses raw textual patent claim data, both supervised and unsupervised state-of-the-art word segmenters, technological dictionaries, entropy measures for wordhood detection, and most importantly an automatic approach to build large reliable lexicons. We show how much each resource contributes to finding the best segmentation.

1. Introduction

This paper attempts to find an optimal solution for the Chinese patent segmentation by combing different pre-trained models with accessible external resources and avoiding any costly annotation. We are most interested in patent claims, whose automatic analysis including data mining and terminology extraction is of significant economic and scientific importance, especially in the context of rapid accumulation of Chinese patent applications since a decade.

Patents in all languages are notoriously rich in new terminology, and inside a patent, the claims, the legally binding part of the patent, contain an even denser terminology than the patent description or the patent abstract. Moreover, the obligation to express each claim in one single sentence makes the structure very different from standard language and particularly hard to analyze. Using syntactic analyzers to process the raw textual patent claim data has been proven helpful in the task of terminology extraction from patents (Yang and Soo 2012). Chinese, however, is a *scriptua continua*, and therefore in a general NLP pipeline, before syntactic parsing and all other kinds of downstream tasks, the chain of characters should be cut into tokens. This is not a problem for common texts with f-scores usually beyond 97% (Zhao and al. 2017). However, if we apply an out-of-the-shelf Chinese word segmenter on patent claims, we have a high percentage of words the segmenter has never seen, so-called Out-of-Vocabulary (OOV) words that considerably degrade the results. All the more tricky but also fascinating from a term extraction perspective is the high percentage of the OOV that are newly created terms, which have not been recorded in any dictionary yet.

Different from many previous works on domain adaptation of Chinese word segmentation, instead of training a new system on abundant annotated data hard to be updated and thus unsuitable for such a domain like patent application changing with each passing day, our work concentrates on evaluating the adaptability and extensibility of general segmenters. The successful adaptation of general segmenters can avoid time-consuming manual corpus labeling works to train a specialized model on specific domain every time. In our experiments, we also verify several hypotheses derived from the data: (1) covering the OOV terms with massive custom dictionaries may help to improve the results; (2) the quality of segmentation may vary between IPC classes; (3) the unsupervised method may have a better performance on the domain-specialized segmentation.

After the brief presentation of linguistic specificities of Chinese language and patent claims, we analyze current difficulties in the word segmentation on Chinese patent claims in section 2. Then, we introduce in section 3 our methods of construction of the external lexicons that are used later in experiments, and in section 4 our annotation framework on test dataset. In the last section, we show the final results and an ablation study on the improvement.

2. Chinese Word Segmentation in specialized domains

As a *scriptua continua* and an isolating language, unlike many alphabetic languages such as French and English, Chinese does not have naturally recognizable linguistic units within written sentences, namely "words". Instead, the sentence in Chinese is a continuous series of characters containing neither white space nor any kind of distinguishable word boundary markers. Only based on this fact, can we understand the long-lasting debate around *wordhood* in Chinese.

In this section, we discuss the wordhood in Chinese from both a linguistic and a technological perspective and show with examples where reside problems of adaptation of segmenters to specialized technological domains.

2.1. Wordhood in Chinese

	咖啡 **ka-fei**	一个 **yi-ge**	小朋友们 **xiao-peng-you-men**
gloss	(transliterated)	one -quantifier	little-friend-friend-plural
meaning	*"coffee"*	*"one" "a(n)"*	*"children"*
GB	咖啡	一 个	小朋友 们
UD	咖啡	一 个	小朋友们
LTP/ THU/ JIE	咖啡	一个 / 一 个	小朋友们 / 小 朋友 们 / …

Tab. 1 – *Examples of the incoherence in Chinese word segmentation between different corpora and standards. GB is the GuoBiao standards[1], UD stands for the Universal Dependencies treebanks[2], and LTP/THU/JIE stand for three state-of-the-art segmenters[3] that are also used later in our*

1 GB/T 13715-1992 *Contemporary Chinese language word segmentation specification for information processing* (《信息处理用现代汉语分词规范》) http://openstd.samr.gov.cn/bzgk/gb/newGbInfo?hcno=B48FFFB924DF90488FEBCB89B91C8869

2 https://universaldependencies.org

3 LTP stands for the segmenter in Language Technology Platform (LTP) (https://www.ltp-cloud.com/), a set of online learning toolkits developed by the Harbin Institute of Technology.

experiments (see section 5).

From the indiscriminate application of western linguistic notion to the most radical opposite that "Chinese does not have words, but instead has characters" (Hoosain 1992; Xu 1997; Packard 2000), currently no common agreement on the definition of "words" in Chinese has been reached. With the rapid development of information technology, the information processing on Chinese language faces a dilemma: While most of the popular tools that are originally developed for western languages require word breaking, the heavy manual process and low inter-annotator agreement make it hard to provide high-quality input corpora to downstream tasks.

In Section 1.1, we investigate the incoherence of existing segmentation criteria: While terms such as "咖啡 ka-fei" have only one possible segmentation, "一个 yi-ge" and "小朋友们 xiao-peng-you-men" are more ambiguous cases in which all of these alternatives should be considered correct. In fact, without absolute standards, the required segmentation largely depends on these downstream tasks. Therefore, the segmentation standards lack practical meaning when the final application objective has not been fully considered.

2.2. Chinese word segmentation for patent claims

The Chinese Word Segmentation (CSW) is very often considered as the first step of various NLP tasks on Chinese, and thus it has unavoidable effects on all kinds of downstream tasks. Comparing to general texts, with their complexity in style and the high percentage of OOV, the patent claims can suffer from even greater noise introduced in segmentation (Fig. 1).

THU stands for the segmenter in THU Lexical Analyzer for Chinese (THULAC) (http://thulac.thunlp.org/), a Chinese NLP toolkit released by Thuhua University.

JIE stands for jiebe (Chinese for "to stutter") segmenter, an individual-developed Chinese Word Segmenter (https://github.com/fxsjy/jieba).

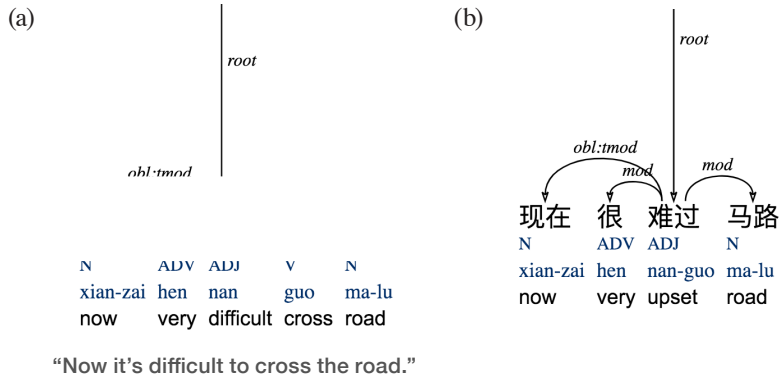


FIG. 1 – *Syntactic parsing result of a Chinese sentence. On the right (b) is the original dependence tree provided by our annotators and visualized by Arborator*[4]*containing an error caused by the wrong segmentation of the term* 难过 *(nan-guo 'sad'), and on the left (a) is the manually corrected version of the same sentence.*

Two main streams of segmentation algorithms are the lexicon-based (Chen et al. 1999 ; Nie, Jin and Hannan 1994) and statistics-based methods. At present most of the popular Chinese segmenters belong to the later that regard the segmentation task as a continuous sequence labeling problem (Ng and Low 2004 ; Low and al. 2005).

Research by Huang (2006) demonstrates that the segmentation errors caused by OOV are in general five times more important than in other cases. Theoretically, once the lexicon has fully covered the vocabulary of the dataset to segment, there should remain few errors concerning only the true ambiguity in original sentences. It is then a natural thought to use domain-specialized lexicons to improve the segmenter's performance on terms never seen in its training dataset. To construct the external dictionary, the simplest and most frequently used resource is the technological dictionaries (Zhao et al. 2010). However, the available dictionaries are unbalanced in terms of domains and often do not include the latest technologies (Rong 2015). Also, regarding the huge quantity and the considerable update speed of terminology in patent applications, a better way is to extract term lists directly from newly

4 Visualization taken from the Arborator annotation tool, Gerdes 2013, https ://arborator.ilpga.fr/

published patent applications and research papers carrying the most recent technological terms, and combine them into the pre-trained model.

In addition to their lexical specificities on OOV terms, at sentence level, patent claims are semi-structured texts with legalese expressions and extremely long sentences very often containing more than 100 characters compared to 20 or 30 characters in general texts. The unusual length of claim sentences, which increases the computing difficulty, is another cause of the low segmentation accuracy on patent claims.

> 3 . 如 权利 要求 2 所述 的 一种 荧光 定量 PCR 检测 鼠疫 耶尔森 氏菌 的 方法 ， 其 特征 在于 ：荧光 定量 PCR 标准 曲线 采用 以 下 步骤 制得 ： 取 不同 DNA 载量 的 质粒 参考 品 各 2 μ l ， 按 上述 荧光 定量 PCR 的 反应 体系 及 反应 程序 在 实施 荧光 定量 PCR 仪|上 进行 扩增 ； 反应 结束 后 根据 得到 的 各个 浓度 的 循 环 阈值 C (t) ， 采用 计算机 自动 绘制 荧光 定量 PCR 标准 曲线 。
>
> *3. The method of **claim** 2, a fluorescent quantitative PCR detection of **Yersinia** pestis claim, wherein: quantitative PCR standard curve was prepared using the following steps: take different amounts of plasmid DNA contained in each of the two reference products μ l, the above-described quantitative PCR reaction system and the reaction procedure in an amplification on the PCR system; after completion of the reaction according to the cyclic threshold value C for each concentration obtained in (t), using the computer to automatically draw the quantitative PCR standard curve.*

The example above is one claim from an actual patent application on Yersinia pestis detecting[5], segmented by the Jieba segmenter (See section 5). The whole claim sentence in Chinese is composed of 137 characters (more than 88 words in English). The underlined character sequences in the example above are where the segmentation errors are found: Even "权利要求 qian-li-yao-qiu", one of the most frequent legalese term in patent texts meaning "claim" is wrongly segmented into "权利 (right)" and "要求 (requirement)". The medical term "耶尔森氏菌 (Yersinia)" and "PCR仪 上 (on the polymerase chain reactor)" are other two segmentation errors in the example sentence: the former is segmented at a position where it should not be, the latter, in contrast, is not correctly segmented where it should be. Typical segmentation errors in this example sentence reveal the difficulty of patent claim processing.

5 Yersinia pestis is the plague bacterium. Patent CN102146467A applied by the Zhejiang Interna-tional Tourism Healthcare Center in 2001.

Construction of lexicons

	Domain	WIKI	CNKI	ELeVe
A	Human Necessities	34 181	9 965	4 157
B	Performing Operations ; Transporting	25 681	9 182	4 226
C	Chemistry ; Metallurgy	27 379	6 820	3 030
D	Textiles ; Paper	10 747	4 841	2 705
E	Fixed Constructions	14 421	8 751	3 613
F	Mechanical Engineering ; Lighting ; Heating ; Weapons ; Blasting	14 308	2 929	3 814
G	Physics	36 281	10 013	3 524
H	Electricity	23 823	4 861	3 824

Tab. 2 – *The International Patent Classification (IPC), established by the Strasbourg Agreement 1971, provides a hierarchical system of language independent symbols for the classification of patents and utility models according to the different areas of technology to which they pertain. In column WIKI, CNKI and ELeVe, we show the total number of terms extracted from the three resources.*

According to WIPO, patents are classified by the IPC standard in technological domains and subdomains. The fundamental distinction is the 8 classes from A to H (Tab. 2). We developed for each IPC classes a custom dictionary, which concatenates three distinct resources into one big word list : We took 1. all page titles from Chinese Wikipedia, 2. keywords used for classifying the academic papers on CNKI.net, and 3. a list of highly autonomous words that was produced as follows : The ELeVe algorithm (Magistry and Sagot 2012) analyzes raw textual data and computes the entropy after each character, i.e. the degree of freedom that the preceding characters offer for the next character.

3. Wikipedia page titles

As the largest online encyclopedia, Wikipedia allows free download of its dump dataset[6]. The zip file downloaded includes the titles of all Chinese Wiki pages (one page title per line and 5 420 881 terms in total in June 2019).

We finally extracted 18 681 Wikipedia page titles in total, divided into 8 domains – note that many titles are long and do not bother the segmentation as long as the precise string is not encountered, and to keep the word lists shorter, we kept only strings that actually appear in the patent domain data of each IPC class by overlapping it with our raw patent application texts and conserving only the termes appearing in the corpus.

3.1. CNKI document keywords

CNKI.net (China National Knowledge Infrastructure, 中国知网) is a Chinese key national information construction project to build and maintain a comprehensive Integrated Knowledge Resources System, including journals, doctoral dissertations, masters' theses, proceedings, newspapers, yearbooks, statistical yearbooks, ebooks, patents, standards, etc. For each of the 8 IPC domains, we construct an article keywords collection by web crawling academical article pages on the site. The classification is based on domain tags on CNKI.net. The obtained keywords are then concatenated to the former Wiki title lists. And again by filtering terms never seen in the patent corpus we reduce the term lists to a reasonable size. The numbers of terms extracted from Wikipedia and from CNKI.net is shown in Tab. 2.

3.2. Lists of term candidates produced by ELeVe

The unsupervised language-independent tokenizer - ELeVe (Magistry and Sagot 2012) is based on the computation of autonomy scores of multi-character terms by measuring the entropy between the characters. The idea of entropy-based segmentation is that a high entropy point is a good potential position for word segmentation, in particular, if the analysis is done bidirectionally. We make use of the ELeVe not only as a segmentation tool (section 5) but also as a resource to produce lists of highly probable terms.

For each IPC class, the segmenter thus provides us with a list of potential words that can be sorted by their degree of autonomy. However, in order to

6 https://dumps.wikimedia.org/zhwiki/

establish a reliable term lists from this raw list, which contains both good terms and bed terms (strings of characters that are not words), we built an automatic perceptron binary classifier model trained on positive and negative examples, selected from the top and the bottom of the list and filtered manually.

From the list of potential words in each IPC domain, we manually selected 2000 positive examples and 2000 negative examples. We also trained a Word-2Vec representation (Mikolov et al. 2013) on our raw corpus by using individual characters as neighbors of the potential words. This allowed us to add to each potential word the vector representations of its 10 closest potential words in terms of their distribution, something that resembles a list of synonyms. For each of these "synonyms", we also provided their degree of autonomy. We gave these 4000 words with different feature combinations of the word itself and for their potential synonyms to the perceptron, which thus trained a Chinese term discriminator model in order to distinguish good from bad term candidates with its best score close to 95%. Features given in training include:

- VecSyn, the vector representations of the 10 most close synonyms of the term
- VecOwn, the vector representation of the term itself
- DistSyn, the distances between the term and its 10 most close synonyms
- AutoSyn, the autonomy scores of the 10 most close synonyms

By means of an ablation study, we obtain the contribution of each feature, shown in Tab. 3. In F-measure results, the perceptron provided only with the vector representation of the term itself and the distances between the term and its 10 most close synonyms has the best accuracy. And the feature that contributes the most is the vector representation of the term itself (accuracy of 0.92 when given only the vector representation of the term itself).

VecSyn	VecOwn	DistSyn	AutoSyn	F-measure
Y	Y	Y	N	0.8892
Y	Y	N	N	0.9304
Y	N	Y	N	0.8636
Y	N	N	N	0.8793
N	N	Y	N	0.8096
N	Y	Y	N	0.9446
N	Y	N	N	0.9219
Y	Y	Y	Y	0.8991

Tab. 3 – *Combinations of different features given to the perceptron have an influence on accuracy. The best result corresponds to the model trained only with VecOwn and DistSyn.*

The high accuracy allows us to dress a large and reliable enough list of potential terms for each of the eight main IPC class with comparatively little manual efforts.

4. Annotation of raw patent claims

With no segmentation evaluation dataset available on Chinese patent claims, we construct our gold test set by randomly selecting a list of 100 lines of claims of each IPC class and segmenting them into segmentation units based on the Guobiao standards[7] and syntactic tests for those terms not found. In his work on customizable segmentation, Wu (2003) distinguishes five types of morphologically derived words in Chinese: 1. Reduplication, 2. Affixation, 3. Directional and resultative compounding, 4. Merging and splitting, 5. Named entities and factoids. We largely followed his classification in our annotation. And since word segmentation can be finer or coarser grained, and we want to compare segmentations that might differ in the granularity of their segmentation, our segmentation markers are annotated with unique symbols of the type of segmentation (Tab. 4): 1.completely syntactically free segmentation units, 2.multi-character expressions (possibly with ambiguous

7 GB/T 13715-1992 *Contemporary Chinese language word segmentation specification for information processing* (《信息处理用现代汉语分词规范》) http://openstd.samr.gov.cn/bzgk/gb/newGbInfo?hcno=B48FFFB924DF90488FEBCB89B91C8869

borders), 3.directional or resultative compounds, 4.classifiers, and 5.morphological affixes.

Marker	Type of cuts	Examples	Translation
space	syntactic unit boundaries	导通 和 断开	conducted and blocked
[]	multi-character expressions	[[中央] [处理器]]	CPU / Central Processor
()	directional/ resultative compounds	(设)有 (接)入	(set up) have_ AUX (link) in/into
|	classifiers	一|种	one | kind
{}	merging/splitting duplication etc. (i.e. idiom)	-	-
_	affixes	超_导体	super_conductor

Tab. 4 – *Distinct symbols used as segmentation markers to indicate six different types of cuts within a sentence. In our patent corpus, we barely found structures in the fifth group.*

We use unique symbols to distinguish six types of possible cuts (Tab. 4). Within all listed types, syntactic unit boundaries indicate the position between two independent syntactic units that without any doubt should be segmented; multi-character expressions markers annotate the inner structure of long terms, mostly noun phrases, which can be segmented on different granularity level, e.g. the term 中央处理器 (zhong-yang-chu-li-qi) can be annotated as a single unit meaning CPU while with a lower granularity we segment it into 中央 (zhong-yang, *"central"*) and 处理器 (chu-li-qi, *"processor"*); the analysis of directional/resultative compounds is reserved to serial verb (or auxiliary) constructions in which the second component denotes some sort of direction or result of the first component; classifiers or measure words are normally necessary between the numeral and the noun in Chinese (e.g. speakers say "one person" or "this person" in English, but "一个人 yi-ge-ren" or "这个人 zhe-ge-ren", here "个 ge" is the classifier); the fifth group is composed of other specific structures in Wu (2003), however, we barely found these structures in patent claims; and the last group is morphological affixes. We marked affixes in our test set only to distinguish them from independent

syntactic units and always regard them as inseparable parts of single words, that is to say, affixes and their radical are never segmented.

5. Experiments and results

We created a corpus of 347 950 claims from the SIPO patent application dataset from November 2017 to April 2018, which, after classification into eight classes, we processed to keep only the patent claims and stripped off all non-Chinese characters. Those characters, including Latin letters, Arabic numbers, punctuation, and all other symbols, are replaced by distinct separation symbols. Moreover, patent claims are highly standardized with specific legalese expressions such as the Chinese equivalents of "we claim", "disclosed is", "the composition of claim 1, wherein". Their number is limited, and they are of no interest for terminology extraction, and we also replaced them with unique symbols as placeholders.

The three supervised segmenters used in the paper are 1. Thulac (Sun et al. 2016), 2. pyltp (Che et al. 2010) and 3. Jieba (https://github.com/fxsjy/jieba). All three of them are state-of-the-art word segmenters frequently used as pre-processors in NLP research projects (e.g. Peng and Dredze 2015, Lizhen et al. 2014, Peng et al. 2017). And all of them accept a list of external terms as a custom dictionary, although they give different priority to the provided terms according to their algorithm.

The experiment results show that giving external word lists to the segmenter does improve the segmentation accuracy (Fig. 2), except for the Jieba segmenter, where we observe no improvements with the dictionary. For the other segmenters, we see that the larger the list, the better the accuracy. In all cases, the externally sourced Wikipedia and CNKI lists give better results than the ELeVe word list. However, the combination of all three lists gives the best results in general. This finding supports the first hypothesis that covering the OOV terms with massive custom dictionaries may help to improve the results.

On the other hand, the unsupervised method does not show a better performance in our experiments compared to supervised ones (Fig. 3). While the LTP and Jieba have no significant gap in segmentation accuracy, the ELeVe is always about 0.2 behind other supervised systems on all granularity levels. But it should be noted that the limitation of memory may prevent ELeVe from taking advantage of its learning ability on enormous data.

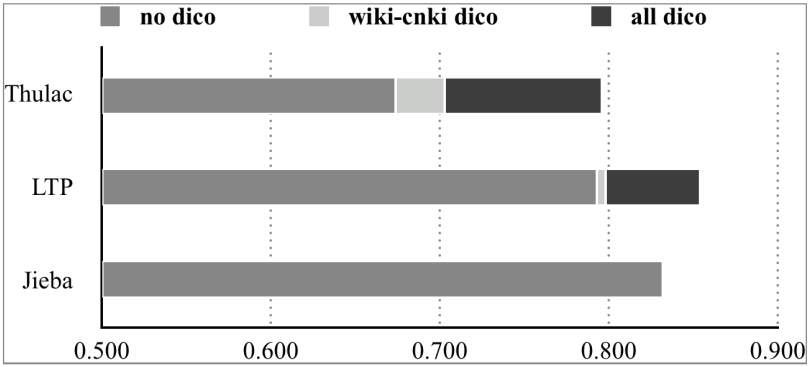


FIG. 2 – *Contribution of external lexicon on supervised segmentation accuracy. The accuracy in this experiment considers all possible cuts as correct segmentation.*

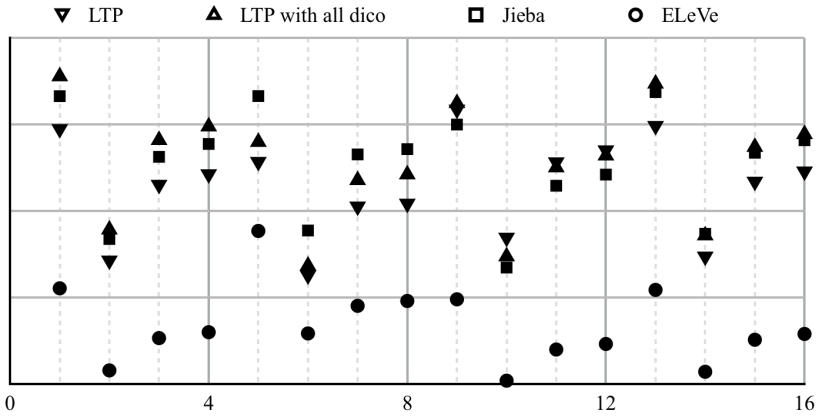


FIG. 3 – *Segmentation accuracy with different segmentation strategies and on different granularity levels. The horizontal axis present the accuracy of segmentation systems. On the vertical axis 1-4 are where we segment only syntactic unit boundaries (spaces); 5-8 segment also all directional/ resultative compounds in addition to all syntactic unit boundaries ; 9-12*

segment also classifiers in addition to all syntactic unit boundaries ; and 13-16 segment all three types above. Multi-character expressions are also segmented according to their granularity level (in each group the granularity level decreases with the growth of the number). For example, the triangle from top left is the accuracy of LTP segmenter with all dictionaries while in the gold file only syntactic unit boundaries (but not spaces inside directional/resultative compounds nor classifiers) are considered as segmentation boundaries as fine the granularity as possible.

In addition, we also observe differences in accuracy between IPC classes (Fig. 4), and as expected these gaps can be reduced with word lists. To investigate if the unsupervised ELeVe segmenter have better performance on larger training datasets, we use the white lines to present the size of dataset for each IPC class. The graph shows no obvious correlation between the size and the accuracy of ELeVe.

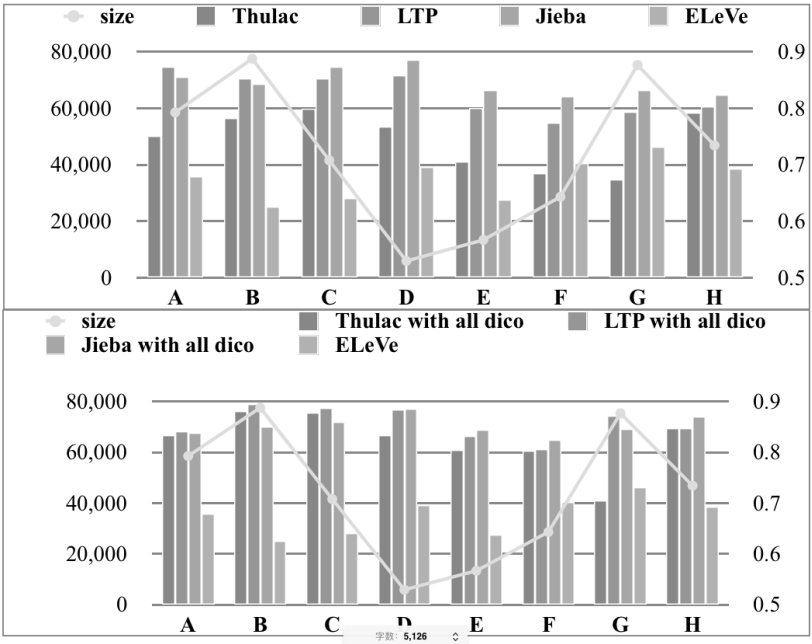


Fig. 4 – *Results of different segments on different IPC classes. As we were interested in the correlation between the performance of segmenters and the size of the corpus, we draw a white line indicating the size of each class.*

The accuracy evaluation on our test set is astonishingly complex but allows nonetheless to draw a few clear conclusions : Accuracy varies widely between 55 % to nearly 90 %, when we vary our parameters (the segmenter, the size of the vocabulary list, the source of the vocabulary list, the patent domain, the size of the required segments, and the types of the required word segmentation).

6. Conclusion

To sum up, the main contributions of the paper are fourfold : Firstly, this is the first work to compare systematically the performance of current segmenters on Chinese patent texts, especially the contribution of the coverage of unknown terms with the help of custom dictionaries extracted from different resources and of different size. Secondly, this is also the first study working on giving a segmentation gold standards to patent claims. Thirdly, the variability of the results shows that an evaluation that does not allow for different granularities of word segmentation cannot compare segmenters in a meaningful way. Lastly, we show how state-of-the-art machine learning techniques can supplement and enhance the extraction of large dictionaries even without a prior word segmentation.

We plan to test these methods again with contextual embeddings, which provide important precision gains on many NLP tasks. Another path is to overcome preliminary segmentation altogether by parsing technological texts with a model that has been trained on character-segmented treebanks. The first results of this method has been presented by Dong et al (2019).

Acknowledgements. We give our gratitude to the China Scholarship Council (CSC) for their financial support.

References

Che, Wanxiang, Zhenghua Li, and Ting Liu. 2010. “Ltp : A chinese language technology platform.” In Proceedings of the 23rd *International Conference on Computational Linguistics* : Demonstrations, pp. 13-16.

Dong, Chuanming, Yixuan Li, and Kim Gerdes. “Character-level Annotation for Chinese Surface-Syntactic Universal Dependencies.” 2019.

Du, Liping, Xiaoge Li, Gen Yu, Chunli Liu, and Rui Liu. 2016. “New word detection based on an improved PMI algorithm for enhancing segmentation

system." *Acta scientiarum naturalium universitatis pekinensis* 52, no. 1 : 35-40.

Gerdes, Kim. "Collaborative dependency annotation." In Proceedings of the second international conference on dependency linguistics (DepLing 2013), pp. 88-97. 2013.

Hoosain, Rumjahn. 1992. "Psychological Reality of the Word in Chinese." In H. C. Chen, & O. J. L. Tzeng (Eds.), *Language Processing in Chinese*, pp. 111-130.

Huang Changning, Zhao Hai. 2007. "Chinese Word Segmentation : A Decade Review." *Journal Of Chinese Information Processing*, 21(3): 8-19.

Liu, Lizhen, Song Wei, Wang Hanshi, Li Chuchu, Lu Jingli. 2014. "A novel feature-based method for sentiment analysis of Chinese product reviews." *China communications*, 11(3), 154-164.

Liu, Yuan, Q. K. Tan, and Xukun Shen. 1994. "Contemporary Chinese Language Word Segmentation Specification for Information Processing and Automatic Word Segmentation Methods."

Low, Jin Kiat, Hwee Tou Ng, and Wenyuan Guo. 2005. "A maximum entropy approach to Chinese word segmentation." In Proceedings of the SIGHAN Workshop on Chinese Language Processing, pp. 448–455.

Lu, Bin, and Benjamin K. Tsou. 2009. "Towards bilingual term extraction in comparable patents." In Proceedings of the 23rd *Pacific Asia Conference on Language, Information and Computation*, Volume 2, vol. 2.

Magistry, Pierre, and Benoît Sagot. 2012. "Unsupervized word segmentation : the case for mandarin chinese." In Proceedings of the 50th Annual Meeting of the *Association for Computational Linguistics* : Short Papers-Volume 2, pp. 383-387.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. "Distributed representations of words and phrases and their compositionality." In *Advances in neural information processing systems*, pp. 3111-3119.

Ng, Hwee Tou and Jin Kiat Low. 2004. "Chinese part-of-speech tagging : One-at-a-time or all-at-once ? word-based or character-based ?" In Conference on Empirical Methods in Natural Language Processing, pp. 277–284.

Packard, Jerome. 2000. *The Morphology of Chinese : A Linguistic and Cognitive Approach*. Cambridge University Press.

Peng, Haiyun, Erik Cambria, & Hussain, A. 2017. "A review of sentiment analysis research in Chinese language." *Cognitive Computation*, 9(4), 423-435.

Peng, Nanyun, Mark Dredze. 2015. "Named entity recognition for chinese social media with jointly trained embeddings." In Proceedings of the 2015 *Conference on Empirical Methods in Natural Language Processing* (pp. 548-554).

Song, Lifeng. 2011. "Research on Chinese Word Segmentation Algorithm for Patent Documents." *Straits Science* 7 : 9-11.

Sun, Maosong, Xinxiong Chen, Kaixu Zhang, Zhipeng Guo, and Zhiyuan Liu. 2016. "Thulac : An efficient lexical analyzer for chinese." *Technical Report*.

Wu, Andi. 2003. "Customizable segmentation of morphologically derived words in Chinese." *International Journal of Computational Linguistics & Chinese Language Processing*, Volume 8, Number 1, February 2003 : Special Issue on Word Formation and Chinese Language Processing 8, no. 1 : 1-28.

Xu, Yi, Alvin Meyer Liberman, and Douglas H. Whalen. 1997. "On the immediacy of phonetic perception." *Psychol. Sci.*, 8, 358–362 . Zanone, P. G. and Kelso, J. A.

Xun, Endong, and Li Cheng. 2009. "Applying terminology definition pattern and multiple features to identify technical new term and its definition." *Journal of Computer Research and Development* 46, no. 1 : 62-69.

Yang, Shih-Yao, and Von-Wun Soo. 2012. "Extract conceptual graphs from plain texts in patent claims." *Engineering Applications of Artificial Intelligence* 25, no. 4 : 874-887.

Yue, Jinyuan, Xu Jin'an, and Zhang Yujie. 2013. "Chinese word segmentation for patent documents." *Acta Scientiarum Naturalium Universitatis Pekinensis* 49, no. 1 : 159-164.

Zhai, Dongsheng, and Ma Wenshan. 2011. "Research the Algorithm of Chinese Patent Claims Segmentation." *Journal of Intelligence* 30, no. 11 : 152-155.

Zhang, Guiping, Liu Dongsheng, Yin Baosheng, et al. 2010. "Research on Chinese Word Segmentation for Patent Documents." *Journal of Chinese Information Processing*, 24(3): 112-116.

Zhang, Rong. 2015. *Terminology and Information Processing*. China Social Sciences Press.

Zhao, Hai, Huang Chang-Ning, Li Mu, and Lu Bao-Liang. 2010. "A Unified Character-based Tagging Framework for Chinese Word Segmentation." ACM Transactions on Asian Language Information Processing, 9(2), pp. 1–32.

Zhao Hai, Cai Deng, Huang Changning, Kit Chunyu. 2017. "Chinese Word Segmentation, a decade review (2007-2017)." The Frontier of Empirical

and Corpus Linguistics, Chunyu Kit and Meijun Liu ed., *China Social Sciences Press*.

Résumé

Cet article a pour objectif de comparer les performances de différents segmenteurs chinois dans des domaines technologiques spécialisés, ainsi que d'évaluer la contribution du lexique externe à leur amélioration. Nous nous intéressons aux textes de brevets dont l'analyse automatique revêt une importance économique et scientifique croissante, et nous tentons de nous attaquer à la source textuelle le plus difficile pour l'extraction de terminologie en termes de langue et de genre : les revendications de brevet chinois. Certains travaux antérieurs sur l'extraction de la terminologie des brevets chinois reposent sur entraînement d'un nouveau modèle ou utilisent des dictionnaires de termes prédéfinis, et aucun ne se concentre sur l'adaptabilité des segmenteurs état-de-l'art existants. Notre approche consiste à utiliser des données textuelles brutes des revendications de brevet, des segmenteurs supervisés et non-supervisés, des dictionnaires technologiques et des mesures d'entropie pour la détection de mots, et surtout une approche automatique permettant de construire de larges lexiques fiables. Nous montrons dans quelle mesure chaque ressource contribue à améliorer la segmentation adaptée.

Analyse des champs lexicaux des acteurs du territoire à partir de corpus textuels sur le web : le cas des controverses autour de l'épandage aérien contre la cercosporiose du bananier en Guadeloupe

Muriel Bonin*, Mathieu Roche*

*CIRAD - Centre de coopération internationale en recherche agronomique pour le développement, Montpellier, France
TETIS, Univ. Montpellier, AgroParisTech, CIRAD, CNRS, IRSTEA, Montpellier, France
muriel.bonin@cirad.fr
mathieu.roche@cirad.fr

Résumé. Les corpus issus du Web sont une source de données qui met en évidence de nouvelles informations dans le cadre des débats publics. Dans cet article, nous proposons une contribution méthodologique fondée sur la fouille de textes afin d'extraire des éléments clés (c'est-à-dire des termes) obtenus à partir de différents acteurs d'un territoire. L'étude de cas sur le traitement aérien contre la cercosporiose en Guadeloupe a été choisie car elle fait l'objet de positions contrastées et opposées. Les résultats montrent des champs lexicaux différents entre acteurs de la société civile (principe de précaution, santé humaine) et groupement de producteur (production). Le chlordécone apparait comme mobilisé par la société civile dans l'argumentaire général qui dénonce un « empoisonnement ». Le « retour de la biodiversité » apparaît récemment dans le discours du groupement de producteurs.

Cet article interroge les apports de l'analyse des traces textuelles issues du web (Kilgarriff & Grefenstette, 2003). Plus précisément, cette étude s'intéresse à l'identification de mondes lexicaux qui peuvent être automatiquement extraits à partir de corpus textuels (Ratinaud & Marchand, 2012). Ces travaux se concentrent sur l'étude du registre lexical lié à la recherche agronomique. Les recherches agronomiques classiques incluent des expérimentations et recueils d'information en laboratoire, en parcelles expérimentales et au champ en condition réelle dans des exploitations agricoles. Les recherches

en sciences sociales mettent en place des enquêtes auprès d'agriculteurs (enquêtes statistiques auprès d'un large échantillon ou échantillon stratifié) et d'acteurs clés du territoire. Les données textuelles issues du web représentent une source d'information qui apporte un éclairage nouveau et complémentaire à certains débats publics. Dans ce contexte, nous proposons une contribution méthodologique qui met en relief des éléments clés de « dires d'acteurs ». Nos propositions ont été mises en œuvre sur une étude de cas concrète qui sera détaillée dans cet article.

Nos travaux cherchent à identifier, de manière semi-automatique, les champs lexicaux de différents acteurs sur un enjeu de territoire. L'analyse a été réalisée en interdisciplinarité sciences sociales (géographie, science politique) et informatique (fouille de textes).

1. Étude de cas : traitements aériens contre la cercosporiose du bananier en Guadeloupe

Nous avons retenu le cas des traitements aériens contre la cercosporiose du bananier en Guadeloupe car ils sont l'objet de prises de position contrastées et opposées de différents acteurs du territoire. Depuis la « crise » du chlordécone (Joly, 2010), les questions environnementales dans la filière bananière sont d'actualité et sont médiatisées, à la fois dans la presse régionale et nationale. Le chlordécone est un insecticide utilisé de 1971 à 1993 aux Antilles françaises pour lutter contre le charançon du bananier (*Cosmopolites sordidus*). Dès la fin des années 1970, Snegaroff (1977) et Kermarrec (1979) mettaient en évidence la contamination des chaînes biologiques en Guadeloupe par les pesticides et les métaux lourds. Une pollution a été « découverte » et saisie par l'État et l'Administration locale à partir du début des années 2000 (Joly, 2010). Suite à une directive européenne en 2009, l'épandage aérien a été l'objet d'une succession d'interdictions puis dérogations entre 2009 et 2014, résultats d'un rapport de force entre société civile et producteurs de banane. Les controverses et prises de positions contrastées de différents acteurs ont été médiatisées et laissent de nombreuses traces sur le web. Ces traces textuelles sont nombreuses et de nature variée. Comment les analyser ? Que peuvent-elles apporter sur les prises de position de différents acteurs ? Nous avons ciblé l'analyse sur cette thématique de l'épandage aérien en production bananière et sur les points de vue contrastés de deux ensembles d'acteurs :

- Des membres de la société civile : les associations de protection de l'environnement et un mouvement social, le LKP (« Liyannaj Kont

Pwofitasyon » en créole, « Collectif contre l'exploitation outrancière » en français), qui a porté une grève générale en 2009 (Daniel, 2009).

- L'Union des Groupements de Producteurs de Bananes de Guadeloupe et Martinique (UGPBAN).

Quels sont les mondes lexicaux de ces acteurs ? L'objectif de cette analyse textuelle est d'éclairer le débat public sur cette question des traitements aériens. Elle est destinée en particulier aux chercheurs en agronomie qui pourront adapter les mises au point de techniques innovantes aux attentes de la société.

2. Grilles d'analyse

Pour répondre à ces questions de recherche, notre cadre d'analyse mobilise des grilles d'analyse en géographie qui placent les acteurs au cœur des territoires et des dynamiques territoriales (Gumuchian *et al*, 2003) et en science politique, sociologie politique : les « 3I » (Idées, Institutions et Intérêts, Palier & Surel, 2005) et les registres de justification (Boltanski, Thevenot, 1991 ; Boltansky, Chiapello, 1999).

Le groupement des producteurs de banane est un acteur incontournable pour ce qui concerne la production bananière. Nous avons retenu également des acteurs de la société civile. En effet, en Guadeloupe, Daniel (2009) montre un retrait du personnel politique traditionnel et une montée en puissance de la société civile dans les Antilles Françaises avec la crise sociale de 2009. Le projet du LKP, qui a porté la grève générale de 2009 est de favoriser à la fois la consommation de produits locaux et la répartition des richesses au profit des guadeloupéens (Ganem, 2010). Le mouvement social de 2009 s'est traduit par des manifestations et barrages et une grève générale de 44 jours. Il a eu des impacts sur l'économie avec une paralysie de l'économie classique et le maintien et développement d'une économie « vivrière » avec la mise en place de marchés improvisés au bord de route dans toute la Guadeloupe, organisés par les agriculteurs avec l'aide du LKP. Le mouvement a aussi eu des impacts sur les consciences avec un débat sur les questions de race et de classe qui sous-tendaient le mouvement. Une solidarité au sein de la population et la valorisation de la culture créole ont été mises en avant (Ganem, 2010). L'accord Bino a été signé avec la mise en place d'une « prime de vie chère » d'un montant de 200€ net par mois sur les salaires de 1,4 SMIC ou moins. La société civile acquiert donc un poids important en Guadeloupe et elle a joué un rôle central dans l'interdiction des traitements aériens en production bananière.

3. Méthode

Le corpus de texte étudié a été constitué à partir d'une requête Google avec les mots-clés suivant : « épandage aérien Guadeloupe LKP associations protection environnement » puis « épandage aérien Guadeloupe UGPAN ». Les liens ainsi trouvés sont nombreux et renvoient principalement à des articles de presse. Notre question de recherche étant d'analyser les champs lexicaux des acteurs, nous avons retenu uniquement les textes écrits par les acteurs eux-mêmes (blogs, communiqués de presse, tracts mis en ligne…). Le corpus ainsi constitué a été analysé au regard des termes extraits avec le logiciel BioTex (Lossio-Ventura *et al*., 2016) qui a été adapté à notre étude de cas. Le nombre de mots des différents corpus est présenté dans le tableau 1. Une description plus complète de ce corpus est donnée dans (Bonin & Roche, 2018).

	Avant 2014	**Après 2014**	**Total**
Société Civile	15117	2504	17621
Groupement Producteurs	8038	10824	18862
Total	23155	13328	36483

Tableau 1 : Nombre de mots des corpus textuels

Les méthodes d'extraction de la terminologie peuvent être guidées par les données (Dobrov & Loukachevitch, 2011) ou par consensus avec les experts (Laporte *et al*., 2012). Notre étude se positionne sur cette première famille d'approche en combinant des méthodes statistiques (Camacho-Collados *et al*. 2014) et linguistiques (Bourigault & Jacquemin, 1999) de manière similaire aux travaux de (Frantzi *et al*., 2000 ; Daille, 1994, etc.). Nos travaux décrits dans le paragraphe suivant consistent à associer une pondération statistique aux termes simples et composés extraits sans prétraitement préalable contrairement aux travaux de (Daille, 1994) qui transforment les syntagmes nominaux en termes binaires avant d'appliquer des mesures statistiques entre éléments composant les syntagmes.

Le logiciel BioTex que nous utilisons dans ces travaux exploite à la fois des informations *statistiques* et *linguistiques* pour extraire la terminologie à partir de textes libres. Cet outil a un caractère générique qui a déjà été exploité pour fouiller des textes du domaine agronomique (Roche *et al*. 2015 ; Lossio-Ventura *et al*., 2016).

Avec BioTex, les informations statistiques apportent une pondération des termes candidats extraits. Cependant, la fréquence d'un terme n'est pas nécessairement un critère de sélection adapté. À titre d'exemple, le mot *épandage* issu de notre requête est présent dans de très nombreux textes. Il n'est donc pas suffisamment discriminant au regard de notre domaine d'étude. Dans ce contexte, des mesures de discriminance et d'autres méthodes de pondérations qui calculent, par exemple, la dépendance des mots composant les termes complexes, peuvent être appliquées.

Mesure de discriminance

Pour effectuer une telle sélection, nos travaux s'appuient sur la mesure TF-IDF. Cette dernière donne un poids plus important aux termes caractéristiques d'un document (Salton & McGill, 1983). Pour attribuer un poids de TF-IDF, il est nécessaire, dans un premier temps, de calculer la fréquence d'un terme (*Term Frequency*). Ainsi, pour le document d_j et le terme t_i, la fréquence du terme dans le document est donnée par l'équation suivante :

$$TF_{ij} = \frac{n_{ij}}{\sum_k n_{kj}}$$

où n_{ij} est le nombre d'occurrences du terme t_i dans d_j. Le dénominateur correspond au nombre d'occurrences de tous les termes dans le document d_j.

La fréquence inverse de document (*Inverse Document Frequency*) mesure l'importance du terme dans l'ensemble du corpus. Elle consiste à calculer le logarithme de l'inverse de la proportion de documents du corpus qui contiennent le terme et est définie de la manière suivante :

$$IDF_i = \log_2 \frac{|D|}{|d_j : t_i \in d_j|}$$

où $|D|$ représente le nombre total de documents dans le corpus et $|d_j : t_i \hat{I} d_j|$ représente le nombre de documents où le terme t_i apparaît. Enfin, la pondération finale s'obtient en multipliant les deux mesures :

$$TF - IDF_{ij} = TF_{ij} \times IDF_i$$

3.1. Mesures d'associations entre les mots

BioTex prend en compte deux facteurs pour extraire la terminologie. Tout d'abord, l'approche extrait des termes selon des patrons syntaxiques définis

(nom-adjectif, adjectif-nom, nom-préposition-nom, etc.). Après un tel filtrage linguistique, un autre filtrage statistique est appliqué. Celui-ci mesure l'association entre les mots composant un terme (par exemple, *agriculture familiale*) en utilisant une mesure appelée C-value (Frantzi *et al*., 2000). Le but de C-value est d'améliorer l'extraction des termes composés particulièrement adaptés pour les domaines de spécialité. Le critère mis en place permet de favoriser les termes n'apparaissant pas, de manière significative, dans des termes plus longs.

Notons que le logiciel BioTex (version en ligne) propose 2 types d'extraction (1200 termes extraits au maximum) : (1) termes composés uniquement, (2) termes simples et composés. Nous avons testé différents paramètres qui apportent des résultats complémentaires (mot ou groupe de mots, longueur des associations de mots, discriminance). Les données produites dans le cadre de ces travaux, à savoir le corpus (Bonin & Roche, 2018) et les termes extraits (Roche & Bonin, 2018), sont en libre accès (https ://dataverse.cirad.fr/dataverse/tetis).

4. Résultats et discussion

4.1. La société civile met en avant le chlordécone et la santé humaine ; le groupement de producteurs la production

Les dix premiers termes extraits (simples et/ou composés) pour Société civile et Groupement de producteurs sont présentés dans le tableau 2. Une liste plus complète de termes extraits ainsi qu'une description des paramètres utilisés sont disponibles dans Dataverse (Roche & Bonin, 2018).

N° de classement	Société Civile	Groupement de producteurs
	Termes simples et composés	
1	Guadeloupe	**Banane**
2	**Chlordécone**	Guadeloupe
3	Environnement	Martinique
4	Épandage aérien	**Bananes**
5	**Santé**	**Filière**
6	État	**Producteurs**
7	Pesticides	**Production**
8	Martinique	**Banane française**
9	Épandage	Agriculture
10	Agriculture	**Banane de Guadeloupe**
	Termes composés	
1	Épandage aérien	**Banane française**
2	**Union Européenne**	**Banane de Guadeloupe**
3	**Principe de précaution**	**Bananes de Guadeloupe**
4	Épandage aérien de pesticides	**Banane durable**
5	**Santé humaine**	**Producteurs de bananes**
6	Aérien de pesticides	**Plan banane durable**
7	**État français**	**Filière banane**
8	**Cancer de la prostate**	Union des groupements de producteurs
9	Interdiction de l'épandage aérien	Groupements de producteurs de banane
10	Interdiction de l'épandage	Groupements de producteurs

Tableau 2 : Termes simples et composés extraits pour Société Civile et Groupement de producteurs et classés selon la mesure C-value.

Les résultats de l'analyse du corpus de texte de la société civile avec terme simple mettent en avant le terme « chlordécone ». Ce résultat est plutôt surprenant étant donné que le traitement aérien est appliqué pour lutter contre la cercosporiose. Le chlordécone était un pesticide utilisé contre le charançon du bananier et interdit en 1993 (mais qui laisse une pollution rémanente). Il s'agit donc d'un problème technique différent. Un retour au corpus de texte montre l'argumentaire mobilisé par la société civile qui fait le lien entre ces deux problèmes techniques différents : l'épandage aérien est associé à un

« l'empoisonnement », de la même manière que le chlordécone. La santé est mise en avant, ainsi que l'État. Notons également que la requête initiale pour la constitution du corpus de texte mentionnait la Guadeloupe seulement. La Martinique apparaît en huitième position dans les termes extraits. Un lien est donc établi par les acteurs de la société civile entre ces deux îles françaises proches géographiquement. Les autres termes extraits sont très généraux par rapport à la thématique étudiée et n'apportent pas d'enseignements particuliers (Guadeloupe, environnement, épandage aérien, pesticides, épandage et agriculture). Seuls les termes en gras sont commentés.

Côté groupement des producteurs, nous constatons que le registre mobilisé met en avant les producteurs et la production (*Banane/Bananes*, *Filière*, *Producteurs*, *Production*). Comme pour la Société Civile, la Martinique est associée à la Guadeloupe et arrive plus tôt dans les termes extraits chez les producteurs (3e terme extrait) que pour la société civile (8e terme). L'origine de la production ressort également dans deux termes composés : *Banane française* et *Banane de Guadeloupe*.

Les termes composés mis en évidence par ce corpus textuel de la société civile sont notamment liés à l'Union Européenne. Les acteurs de la société civile font appel à une directive européenne qui interdit l'épandage aérien. Le principe de précaution et la santé humaine sont mis en avant et précisés par le cancer de la prostate. L'État français apparaît également.

Pour les groupements de producteurs, les termes composés extraits confirment le registre principal utilisé autour de la production et des producteurs.

4.2. Termes (simples et composés) extraits avec discriminance pour Société Civile et Groupement de producteurs

Pour la société civile, le terme *enfant* apparaît dans les termes extraits avec discriminance (tableau 3). Il s'agit ici de mettre l'accent sur les effets particuliers des pesticides sur la santé des enfants. Dans les termes composés pour la société civile, on retrouve le thème des enfants, mais avec ici l'utilisation du créole *(Ti Moun an nou*, « nos enfants à nous »). Outre les termes déjà vu précédemment, on peut noter ici le nom de *Luc Multigner* dont les prises de position ont été très controversées : lanceur d'alerte, il établissait un lien entre le chlordécone et les cancers. Il a été critiqué pour ne pas disposer de données suffisantes pour établir ce lien. Cependant, des études ultérieures, notamment l'étude *Timoun*, ont confirmé les liens entre exposition au chlordécone et can-

cer de la prostate et exposition des mères et développement des jeunes enfants (Boucher *et al*., 2013 ; Costet *et al*., 2015).

Plusieurs sujets de controverses apparaissent chez les groupes de producteurs : la *biodiversité* et *retour de la biodiversité* renvoient aux débats suite aux études montrant que les plantations bananières hébergent une riche biodiversité. Des arguments viennent à l'encontre de cette affirmation en mettant en avant la disparition de biodiversité liée aux plantations bananières. La possibilité d'une diversification de l'agriculture ne fait pas non plus consensus, de même que l'attribution des subventions, les décisions et les alternatives possibles (*agriculture de préservation*, *système de pulvérisation*).

N° de classement	Société Civile	Groupement de producteurs
Termes simples et composés		
1	**Enfants**	Banane française
2	Aides	**Biodiversité**
3	Eaux	Gouvernement
4	Pétition	**Diversification**
5	Prix	Population
6	Agriculture biologique	**Décision**
7	Mêmes	**Agriculture de préservation**
8	Union	**Subventions**
9	Culture	**Système de pulvérisation**
10	Élysée	Lettre ouverte
Termes composés		
1	Principe de précaution	Banane française
2	Union européenne	**Retour de la biodiversité**
3	Collectif contre l'épandage	**Solutions alternatives**
4	Euros d'amendes	Grand sachant
5	**Moun an nou**	Outre-mer
6	Agriculture biologique	République dominicaine
7	Parlement européen	**Agriculture de préservation**
8	Mise sur le marché	**Système de pulvérisation**
9	**Luc Multigner**	**Biodiversité dans les bananeraies**
10	Pesticides agricoles	Lettre ouverte

Tableau 3 : Termes simples et composés extraits et classés selon la pondération fondée sur le TF-IDF (mesure F-TFIDF-C décrite dans (Lossio-Ventura et al. 2016)) pour Société Civile et Groupement de producteurs. Analyse avant et après l'interdiction des traitements aériens (2014)

4.3. Analyse avant et après l'interdiction des traitements aériens

Nous retenons dans cette section seulement les éléments nouveaux par rapport aux analyses antérieures. Remarquons dans le tableau 4 pour la société civile, que les troubles sont mis en avant. Il ne s'agit plus seulement d'appliquer le principe de précaution comme indiqué avant l'interdiction mais plutôt de prendre en charge les troubles sur la santé liés aux pesticides.

Dans les mots composés après 2014, les questions de *kilo de matière fraiche*, *microgrammes par kilo* apparaissent et renvoient à la charge en chlordéone contenue dans les aliments. Ici aussi il s'agit du problème lié à la pollution par la chlordécone, et non aux traitements aériens mais qui sont réunis dans la dénonciation d'une pollution générale par les pesticides.

N° de classement	Société Civile avant 2014	Société Civile après 2014
Termes simples et composés		
1	Union	Chlordécone
2	Aides	Enfants
3	Pétition	Guadeloupe
4	Programme	**Santé**
5	Eaux	**Troubles**
6	Prix	Martinique
7	Agriculture biologique	Environnement
8	Culture	Pesticides
9	Élysée	Antilles
10	Entrée	Pesticide
Termes composés		
1	Union européenne	Antilles françaises
2	Collectif contre l'épandage	Luc Multigner
3	Euros d'amendes	Pesticides agricoles
4	Moun an nou	Sud Basse-Terre
5	Agriculture biologique	**Kilo de matière fraîche**
6	Parlement européen	Docteur Luc Multigner
7	Mise sur le marché	Kilo de matière
8	**Principe de précaution**	**Microgrammes par kilo**
9	Prise en charge	Association envie-santé
10	Victimes du médiator	Docteur Luc

Tableau 4 : Termes simples et composés extraits et classés selon la pondération fondée sur le TF-IDF (mesure F-TFIDF-C décrite dans (Lossio-Ventura et al. 2016)) pour Société Civile avant et après l'interdiction des épandages aériens en 2014.

Le tableau 5 nous montre une évolution dans le champ lexical du groupement de producteurs : la question du retour de la biodiversité est présente après 2014 et pas avant. L'argument biodiversité a été utilisé pour « redorer » l'image de la production bananière d'un point de vue environnemental. Cependant, nous ne pouvons pas conclure que cette évolution soit liée à l'interdiction des traitements aériens. Un autre élément a marqué cette période : la parution des résultats de l'étude Timoun, à partir de 2013, qui ont établi un lien entre l'exposition de la mère au chlordécone et le développement de l'enfant (Boucher *et al*., 2013 ; Costet *et al*., 2015).

N° de classement	Groupement de producteurs avant 2014	Groupement de producteurs après 2014
Termes simples et composés		
1	Gouvernement	Banane
2	Diversification	Bananeraies
3	Population	Espèces
4	Subventions	**Retour de la biodiversité**
5	Système de pulvérisation	Plantations
6	Décision	**Biodiversité dans les**
7	Lettre ouverte	**bananeraies**
8	Soutien public	Rayon
9	Tribunal administratif	Guadeloupe
10	Préservation	Banane antillaise
		Bananeraies de Guadeloupe
Termes composés		
1	Grand sachant	**Retour de la biodiversité**
2	Agriculture de préservation	Banane antillaise
3	Système de pulvérisation	**Biodiversité dans les**
4	Lettre ouverte	**bananeraies**
5	Outre-mer	Bananeraies de Guadeloupe
6	République dominicaine	Bananes de Guadeloupe et
7	Soutien public	Martinique
8	Tribunal administratif	Île de France
9	Secteur de la banane	Antilles françaises
10	Moteur du progrès	Base de banane
		Cahier des charges

Tableau 5 : Termes simples et composés extraits et classés selon la pondération fondée sur le TF-IDF (mesure F-TFIDF-C décrite dans (Lossio-Ventura et al. 2016)) pour Groupement de producteurs avant et après l'interdiction des épandages aériens en 2014.

5. Conclusion

Les résultats montrent des registres différents dans les champs lexicaux des différents acteurs : principe de précaution et santé humaine pour les acteurs de la société civile ; registre de la production pour le groupement de producteurs. Nous avons mis en évidence des nuances dans les champs lexicaux de différents courants de la société civile, ainsi qu'une évolution des champs lexicaux des acteurs de la société civile et du groupement de producteur au fil du temps. Des acteurs « alliés » apparaissent dans les corpus de texte, ce qui permet de poser des hypothèses sur les coalitions de cause (Sabatier *et al*., 1993).

Nos analyses présentent cependant quelques limites. Une première restriction de nos travaux tient au fait que les termes nominaux peuvent véhiculer des concepts différents des termes verbaux voire adjectivaux. Une couverture plus large quant à l'extraction de la terminologie semble tout à fait pertinente à réaliser dans nos futurs travaux. Une seconde limite concerne l'étude qui est ici circonscrite aux données « open », c'est-à-dire en libre accès sur internet. Certains aspects de la problématique ne sont pas mis en évidence : il existait des désaccords entre le LPG (Les producteurs de Guadeloupe) et Banamart (le groupement des producteurs de banane en Martinique) au sujet des traitements aériens. Les textes en ligne en libre accès sur Internet issus de l'union des groupements (UGPBAN) ne rendent pas compte de ces divergences. Par ailleurs, certains acteurs ne s'expriment pas à travers le web. C'est le cas des ouvriers de la banane. Suite à l'interdiction des traitements aériens, deux pratiques sont mises en place pour lutter contre la cercosporiose : l'effeuillage (enlever les feuilles infestées par le champignon pour éviter sa propagation) et les traitements par pompe à dos. Les ouvriers qui appliquent ces traitements sont particulièrement exposés aux pesticides. L'exposition a été déplacée et concentrée sur ces travailleurs. Ce thème est absent des textes que nous avons analysés.

Les enquêtes classiques de terrain restent indispensables pour identifier ces dimensions qui ne sont pas disponibles en « open » sur le web. De manière plus générale, l'analyse des « mondes lexicaux » des différents acteurs sur le web ne remplace ni enquêtes de terrain, ni lecture des textes. Elle apparaît comme complémentaire des méthodologies plus classiques d'enquêtes afin d'identifier des hypothèses pour mieux cibler les guides d'entretiens d'enquêtes de terrain ou conforter sur un large échantillon des hypothèses issues d'enquêtes. Ces analyses ouvrent des perspectives pour assurer une orienta-

tion vers une recherche agronomique en phase avec les attentes de la société et en particulier de la société civile.

Remerciements

Cette étude a été réalisée dans le cadre du projet BigDataPol : http://text-mining.biz/Projects/BigDataPol. Ce projet CRESI (Créativité et innovation scientifiques) du Cirad porté par Jean-François Le Coq vise à mobiliser des approches de *Big Data* pour l'analyse des processus et des effets des politiques publiques dans le milieu rural.

References

Bonin, Muriel and Roche, Mathieu. 2018. "Corpus 'Controverses sur l'épandage aérien en Guadeloupe", doi : 10.18167/DVN1/LSGN42, CIRAD Dataverse

Boucher, Olivier *et al*. 2013. "Exposure to an organochlorine pesticide (chlordecone) and development of 18-month-old infants". In *NeuroToxicology*, 2013, 35, p. 162-8.

Bourigault, Didier and Jacquemin, Christian. 1999. "Term Extraction + Term Clustering : An Integrated Platform for Computer-Aided Terminology". In *Proc. of EACL 1999 (European Chapter of the Association for Computational Linguistics)* : 15-22

Camacho-Collados, José *et al*. 2014. "Approche statistique pour le filtrage terminologique des occurrences de candidats termes en texte intégral". In *Proc. of JADT 2014 - Journées internationales d'Analyse statistique des Données Textuelles*.

Costet, Nathalie *et al*. (2015), "Perinatal exposure to chlordecone and infant growth", in *Environmental Research*, 142, p.123-134

Daille, Béatrice. 1994. "Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques". *Ph.D. Thesis, Univ. Paris 7*

Daniel, Justin. 2009. "La crise sociale aux Antilles françaises. Retour sur la question sociale et reflux du politique", *EchoGeo*, 8p.

Dobrov, Boris and Loukachevitch, Natalia. 2011. "Combining Evidence for Automatic Extraction of Terms". In *Proc. of Patten Recognition and Machine Intelligence, LNCS*, p. 235-241

Franzi, Katerina. 2000. “Automatic recognition of multi-word terms : the C-value/NC-value method”. *Int. Jour. on Digital Libraries*, 3(2), p. 115-130

Ganem, Valérie. 2010. “Retour sur le ’Liyannaj Kont Pwofitasyon (LKP)‘ accompli en Guadeloupe”, in *Nouvelle revue de psychosociologie*, n° 9, p. 199-211.

Gumuchian, Hervé *et al*. 2003. “*Les acteurs, ces oubliés du territoire*”. Anthropos, 186 p.

Joly, Pierre-Benoit. 2010. “*La saga du chlordécone aux Antilles françaises. Reconstruction chronologique 1968-2008*”. Rapport du projet AFSSET action 39 du Plan National Chlordécone 2008-2010, Inra Unité Sens en Sociétés, Paris. Juillet, 82 p.

Kermarec, Alain. 1979. “*Niveau actuel de la contamination des chaînes biologiques en Guadeloupe : pesticides et métaux lourds*”. INRA Guadeloupe/ministère de l’Agriculture, 155 p.

Kilgarriff, Adam and Grefenstette, Gtregory. 2003. “Introduction to the special issue on the web as corpus », in *Computational Linguistics - Special issue on web as corpus archive*, Vol 29, Issue 3, p. 333-347

Laporte, Marie-Angélique, *et al*. 2012. “ThesauForm-Traits : A web based collaborative tool to develop a thesaurus for plant functional diversity research”. *Ecological informatics (Special Issue)*, Vol. 11, p. 34-44

Lossio-Ventura, Juan-Antonio, *et al*. 2016. “Biomedical term extraction : overview and a new methodology”, in *Information Retrieval Journal*, 19(1-2), p. 59-99

Palier, Bruno, Surel, Yves. 2005. “Les ’trois I’ et l’analyse de l’Etat en action ». *Revue française de science politique*, 2005/1, p.7-32

Rangeon, François. 2004. “*Société civile : histoire d’un mot*”. Édition Inclinaison, 52 p.

Roche, Mathieu, Bonin, Muriel. 2018. “*Termes « Controverses et épandage aérien en Guadeloupe »*”, doi :10.18167/DVN1/37ENLP, CIRAD Dataverse

Roche, Mathieu *et al*. 2015. “Extraction automatique des mots-clés à partir de publications scientifiques pour l’indexation et l’ouverture des données en agronomie”. *Cahiers Agricultures*, Volume 24, Numéro 5, p. 313-320

Salton, Gerard, McGill, Michael. 1983. “*Introduction to Modern Information Retrieval*”. *McGraw-Hill*

Snegaroff, Jacques. 1977. “Les résidus d’insecticides organochlorés dans les sols et les rivières de la région bananière de Guadeloupe”. *Phytiatrie-Phytopharmacie*, n° 26, p. 251-268.

Ratinaud, Pierre and Marchand, Pascal. 2012. “Application de la méthode ALCESTE à de « gros » corpus et stabilité des ‘mondes lexicaux’ : analyse du ‘CableGate’ avec IRaMuTeQ”. In : *Actes des 11e Journées internationales d’Analyse statistique des Données Textuelles*, JADT 2012, Liège, p. 835-844.

Abstract

Corpora from the web is a data source that highlights new piece of information in the context of public debates. In this paper, we propose a methodological contribution based on textmining in order to extract key elements (i.e. terms) obtained from different actors of a territory. The case study about aerial treatment against cercosporiose in Guadeloupe was chosen because it is the subject of contrasting and opposing positions. The results highlight different lexical fields between civil society actors (precautionary principle, human health) and producer groups (production). Chlordecone appears as mobilized by civil society in the general argument that denounces a “poisoning”. The “comeback of biodiversity” appears recently in the speech of the producer group.

Analysing clinical trial outcomes in trial registries : towards creating an ontology of clinical trial outcomes

Anna Koroleva*, Corentin Masson**, Patrick Paroubek***

*LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay ; AMC, University of Amsterdam, Amsterdam, Netherlands
koroleva@limsi.fr
** LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay ; AMF (French Financial Market Authority), France
corentin-masson@outlook.fr
*** LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay
pap@limsi.fr

Abstract. A clinical trial is a study that evaluates the effects of one or several interventions on a certain population regarding some outcomes - variables that are monitored to assess the impact of the intervention. Trial outcomes are one of the crucial characteristics of a clinical trial. Outcomes are defined by several aspects, such as the name of the variable monitored, measurement tool used, timepoints, analysis metric, aggregation method. We propose to semi-automatically create a structured database of trial outcomes and aspects defining them, that can be used as support for outcome extraction task or to the development of Core Outcome Sets (COS). We propose to use the data from trial registries – online databases containing information about planned and conducted clinical trials, including outcomes. We apply supervised and unsupervised Natural Language Processing techniques to describe and analyse trial outcomes extracted from registries.

1. Introduction

A clinical trial is a study that evaluates the effects of one or several interventions on a certain population regarding some health-related parameters, called outcomes[1]. Outcomes in clinical trials are variables that are monitored

1 https://www.who.int/topics/clinical_trials/en/

to establish the impact of the explored intervention on the health of the population studied. Trial outcomes are one of the crucial characteristics of a clinical trial as they reflect the research question and the explored hypothesis of a trial.

Outcomes are defined by several dimensions. The description of an outcome always comprises a definition of the variable monitored. It can be numerical (temperature), binary (occurrence of an event), or qualitative (quality of life). Some outcomes can be difficult to measure directly, so various measurement tools can be used, such as questionnaires or scales. Outcomes can be measured objectively or subjectively, recorded by a clinician or patient-reported. An outcome is measured several times during a given trial, and these timepoints should be specified for each outcome.

Various analysis metrics can be used for analysing an outcome at the participant level : change from baseline, final value, time to event. At the group level, outcomes are analyzed using some method of aggregation (mean, median, proportion). For the final analysis of the studied population, two main types of analysis can be used : intention-to-treat analysis[2] (all the patients enrolled are analyzed, including those who dropped out of the trial) and per-protocol analysis[3] (only patients who followed the clinical trial instructions are included into the analysis).

In this paper, we propose to create semi-automatically a database containing information about outcomes used in randomized controlled trials (RCTs), related measurement tools, timepoints, analysis metrics and aggregation methods used. Such a database could be used as support for outcome extraction task (Blake and Lucic, 2015 ; Demner-Fushman *et al*., 2006 ; Blake and Lucic, 2016 ; Summerscales *et al*., 2009) or could contribute to the development of Core Outcome Sets (COS) – agreed standardised sets of outcomes (and related measurement tools) that should be reported for each specific medical domain[4] to facilitate summarising and practical use of research results (Clarke and Williamson, 2016).

The structure of this paper is as follows : first, we describe the data source that we propose to use to build a database of outcomes. After that, we describe the textual features of outcomes in registries, and we report on our first exper-

2 https ://www.nice.org.uk/glossary ?letter=i
3 https ://www.nice.org.uk/Glossary ?letter=P
4 http://www.comet-initiative.org/glossary/cos/

iments on building a structured database of trial outcomes, using unsupervised or semi-supervised clustering to normalize the outcome descriptions.

2. Data

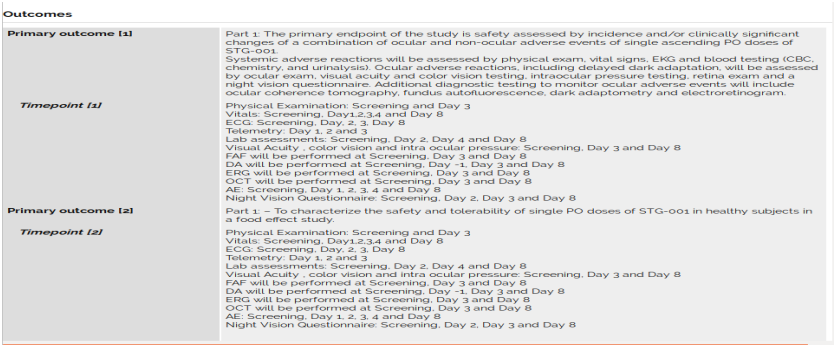
We propose to use the data from trial registries – online databases containing information about planned and conducted clinical trials, such as studied medical condition, treatment(s), population, outcomes, etc. Information in registries is presented in a structured form; all the registries have data fields for outcomes, usually with division into primary (the most important) and secondary outcomes.

Our starting point is a corpus of 3,938 articles from PubMed Central[5] with the publication type "Randomized controlled trial". For 2,701 articles from this corpus, we were able to find the trial registration number in the text using regular expressions. In some texts, there were several registration numbers mentioned (reporting several trials in one paper, referring to previous trials etc.); for some registration numbers, entries were found in several registries. We searched 13 trial registries and the WHO portal. We downloaded and parsed the data from corresponding trial registries for the obtained registration numbers, and we extracted the fields describing primary and secondary outcomes. If the data for the same trial registration number was available in several registries, we downloaded all the versions, since for a given outcome the wording or the structuring of the description may differ. This work resulted in a corpus of 17,515 outcome descriptions (11,182 unique outcome entries).

3. Textual features of outcomes

The level of detail in the outcome field varies. The field can contain only a noun phrase naming the measured variable (e.g. "*body weight*"), or a free-text description of the outcome and related information elements (e.g. "*The outcome of interest was self-reported medication side effects ever up until the time of interview in 1994, and was recorded as Yes or No.*"). The length of outcome descriptions in our dataset ranges from 2 to 6,606 characters (median = 81, mean = 197.4). Shorter outcomes (up to 6 symbols) are often represented by an abbreviation.

5 https://www.ncbi.nlm.nih.gov/pmc/



Outcomes	
Primary outcome [1]	Part 1: The primary endpoint of the study is safety assessed by incidence and/or clinically significant changes of a combination of ocular and non-ocular adverse events of single ascending PO doses of STG-001. Systemic adverse reactions will be assessed by physical exam, vital signs, EKG and blood testing (CBC, chemistry, and urinalysis). Ocular adverse reactions, including delayed dark adaptation, will be assessed by ocular exam, visual acuity and color vision testing, intraocular pressure testing, retina exam and a night vision questionnaire. Additional diagnostic testing to monitor ocular adverse events will include ocular coherence tomography, fundus autofluorescence, dark adaptometry and electroretinogram.
Timepoint [1]	Physical Examination: Screening and Day 3 Vitals: Screening, Day1,2,3,4 and Day 8 ECG: Screening, Day, 2, 3, Day 8 Telemetry: Day 1, 2 and 3 Lab assessments: Screening, Day 2, Day 4 and Day 8 Visual Acuity , color vision and intra ocular pressure: Screening, Day 3 and Day 8 FAF will be performed at Screening, Day 3 and Day 8 DA will be performed at Screening, Day -1, Day 3 and Day 8 ERG will be performed at Screening, Day 3 and Day 8 OCT will be performed at Screening, Day 3 and Day 8 AE: Screening, Day 1, 2, 3, 4 and Day 8 Night Vision Questionnaire: Screening, Day 2, Day 3 and Day 8
Primary outcome [2]	Part 1: – To characterize the safety and tolerability of single PO doses of STG-001 in healthy subjects in a food effect study.
Timepoint [2]	Physical Examination: Screening and Day 3 Vitals: Screening, Day1,2,3,4 and Day 8 ECG: Screening, Day, 2, 3, Day 8 Telemetry: Day 1, 2 and 3 Lab assessments: Screening, Day 2, Day 4 and Day 8 Visual Acuity , color vision and intra ocular pressure: Screening, Day 3 and Day 8 FAF will be performed at Screening, Day 3 and Day 8 DA will be performed at Screening, Day -1, Day 3 and Day 8 ERG will be performed at Screening, Day 3 and Day 8 OCT will be performed at Screening, Day 3 and Day 8 AE: Screening, Day 1, 2, 3, 4 and Day 8 Night Vision Questionnaire: Screening, Day 2, Day 3 and Day 8

Figure 1 : An outcome entry

Structure of registries differs. Some registries have a separate field for each of trial outcomes, others have only one field where a list of outcomes is recorded. Each item of the list can contain several sentences, describing all the outcome-related information. Some registries have separate fields for outcome timepoints or for outcome measurement tools, while in others all the outcome-related information is recorded in one field.

Figure 1 shown an example of an outcome entry from the Australian New Zealand Clinical Trials registry.

4. Methods

We propose to use Natural Language Processing (NLP) techniques (rules, deep learning and clustering methods) to create a database with structured information on outcomes, based on data extracted from trial registries. We address normalisation of primary outcomes extracted from trial registries and extracting related information.

4.1. Clustering

To assess the variability of outcome descriptions, we used unsupervised clustering. There are several methods of clustering :

1. *Content Mapping methods* : transformation of words to concepts extracted from ontologies (WordNet) to obtain a vector containing each concept representing every document (outcome entry). The vectors can be analysed using Bag-of-words and TF-IDF approaches.

Singular value decomposition (SVD) can be applied to reduce the dimensionality to improve clustering with K-means or hierarchical agglomerative clustering (HAC) (cf. Termier *et al*., 2001). The clustering algorithm can be modified to change the used disance (cosinus, euclidian) to graph distances like Wu-Palmer so that the algorithm can exploit semantic distance to identify clusters.

2. *Word embeddings methods*: language models trained with neural networks, such as word2vec, can be used to obtain word embeddings without using ontologies.
3. *Hybrid methods*: combining classic and word embeddings methods

4.2. Rules

To normalise an outcome entry, we first need to determine if a description contains one outcome or a list of outcomes. Normalising single short outcome descriptions is a rather simple task for which we perform abbreviation expansion by simple regular-expression-based approach using the text of the article related to a registry entry to search for possible expansions of abbreviations.

Lists of outcomes should be divided into single outcomes. It should be taken into account that a list may be present within a description of a single outcome, e.g. a list of measurement tools used, which should not be separated at this stage.

Items describing an outcome may be defined in several sentences, e.g.:

The primary outcome is the change in child problem behavior after intervention. The following instruments will be applied: 1. Strengths and Difficulties Questionnaire (SDQ); 2. Eyberg Child Behaviour Inventory (ECBI).

Although such cases are more difficult for analysis than one-sentence entries, the number of constructions used to describe an outcome and related information elements is limited, allowing to create a set of rules to extract the information.

4.3. Supervised machine learning

Supervised machine learning can be used to extract information from long free-text outcome descriptions, using an annotated corpus to train. For this goal, we annotated 2000 sentences for mentions of primary outcomes, e.g. (outcome is in bold):

The primary outcome was ***the change from baseline in airway resistance (sRaw) at 12 hrs post dose measured by whole body plethysmography***.

We annotated text spans containing all the outcome-related information (outcome and measurement tool name, timepoints, etc.).

We focused on primary outcomes in our annotation efforts as they are the most important information element for our main goal of outcome switching / spin detection (Koroleva and Paroubek, 2018).

Our experiments on applying supervised machine learning to the outcome extraction task are described in detail elsewhere. We compared several approaches and models to choose the best performing method. In brief, the chosen method, proposed by Devlin *et al*. (2018), consists in pre-training deep bi-directional language representations on a large unannotated corpus and consequently fine-tuning them on a rather small annotated corpus for a supervised task. We compared a number of language models, including BERT (Devlin *et al*., 2018), BioBERT (Lee *et al*., 2019) and SciBERT (Beltagy *et al*., 2019).

5. Results

5.1. Clustering

The first experiment was based on Content Mapping Method. Using POS-Tag techniques, disambiguation and WordNet, we transform each outcome into a list of synsets. In the first approach, we mapped those synsets into vectors using TF-IDF ; in the second approach, we mapped this TF-IDF into a smaller matrix using SVD. Results are not satisfying. As expected, intra-cluster variance is decreasing with the number of clusters (cf. Table 1 and 2), but there is no significant drop that would allow us to select an optimal number of clusters.

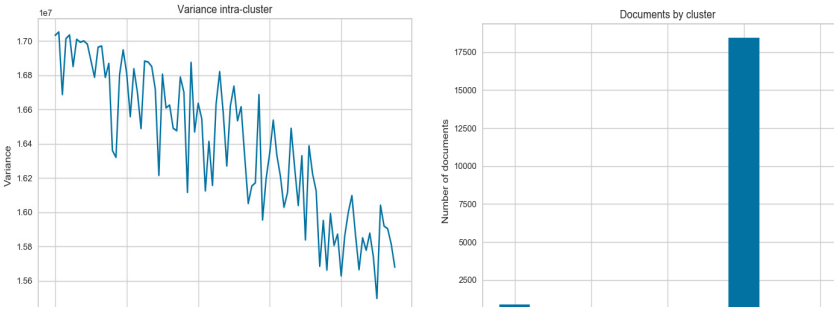


Table 1 : TFIDF : Variance depending on number of clusters ; cluster sizes

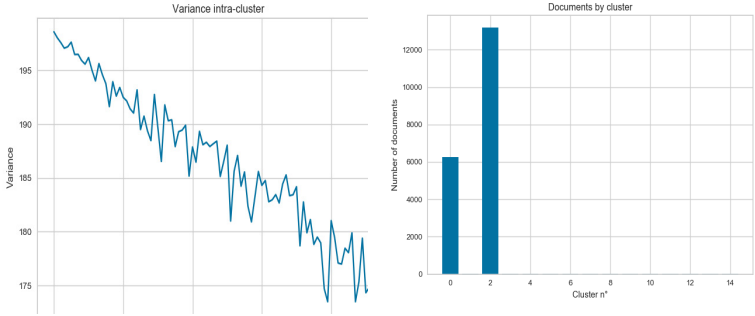


Table 2 : SVD : Variance depending on number of clusters ; cluster sizes

In our second experiment, we tried to add hypernyms into token vectors to improve the results. Adding those tokens, we hoped to add generality and see clusters merging as we go from hypernym to hypernym. We tried to keep only the first hypernyms along original synsets, or to keep everything, getting a large token vector. Figure 2 shows intra-cluster variance as we add more clusters in K-means. Each curve represent one more hypernym taken into account. The variance decreases as we take more hypernyms. We still can not find an optimal number of clusters.

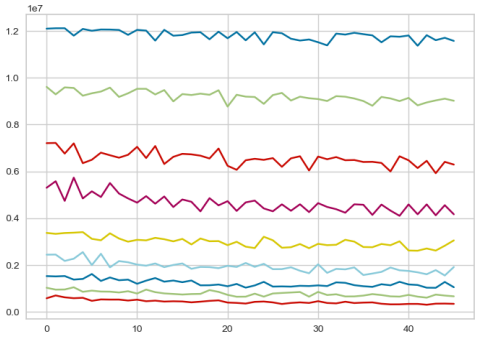


Figure 2 : Variance when including hypernyms in vectors

5.2. Supervised machine learning and rules

The best performing deep-learning model (BioBERT fine-tuned for primary outcome extraction) showed the precision of 86.99 %, recall of 90.07 % and F-measure of 88.42 %.

We suggest to use the deep learning algorithm to extract outcome mentions (such as "*the change from baseline in airway resistance (sRaw) at 12 hrs post dose measured by whole body plethysmography*") and to consequently use simple pattern-based rules to extract outcome-related information. For example, measurement tool name can be extracted using a regular expression pattern "*(.*)\s*,?\s*(?:as|which \w+|that \w+)? (?:measured|assessed|defined|rated|quantified|marked|tested|recorded) (?:with|by|as|using|on|through) (.*)*". Timepoints can be extracted as prepositional phrases containing words with semantics of time (*day*, *month*, *baseline*, etc.). Aggregation method, analysis metrics and type of analysis can be extracted using a dictionary of relevant words (*mean*, *change*, *per-protocol*, etc.).

6. Discussion

6.1. Issues encountered

Ones of the encountered difficulties consists in separating coordinated outcomes (e.g. "BMI and aerobic fitness"). Syntactic analysis of such phrases

to identify coordinated entities is not likely to be useful because of common errors in parsing incomplete sentences. The task is further complicated by possible presence of coordination within one outcome, which does not need to be divided, and by the need for ellipsis analysis to obtain correct outcome names for some cases (e.g. "local and regional control" which would need to be divided into "local control" and "regional control"). At the current stage, we have not resolved this issue.

Another important issue raised by this work is the absence of uniformity of describing outcomes in registries regarding the length, the details included, and the structure of descriptions (free text vs. noun phrases).

We faced some problems during our clustering experiments. First, outcomes are often represented by short texts containing a high proportion of specific words. Our approach based on WordNet is not efficient for domain-specific documents because a high percentage of words are not present in WordNet. A way to solve this issue would be to use a biological ontology or switch to word embedding methods, potentially using BioBERT representations.

For the experiments using hypernyms to generalize a document, the problem we faced is whether to keep all levels of hypernyms or not. Each word in a document being at a different depth of the WordNet, iteratively taking hypernyms for all words does not result in the same level of generality for each original synset. A word being level deeper than another one will not merge using this technique, thus we have to select the optimal hypernym for each word. TF-IDF & cosine similarity do not give better results. We should try to use Wu-Palmer as distance for the clustering algorithm.

6.2. Future work

The current experiments get its inspiration from the Lesk Algorithm and the paper of Scheepers *et al*. (2018). The idea behind it is to be able to extract the good level of hypernyms without adding noise to the document. For this end, we take the definition of each extracted synset, which will represent sense of the word. For each level of hypernym until reaching the root, we measure the distance between the definition of the current hypernym with the one of the original synset. We take into account each hypernym until the distance goes beyond a certain threshold. We can choose a distance measure, based on Wordnet (Wu-Palmer similarity) or based on word-embeddings (e.g. Glove, Paragram, Bert, Elmo). When the procedure is accomplished, we should have

a generalized document that might be more adequate to clustering, using TF-IDF, SVD or even on word-embeddings.

7. Conclusion

In this paper, we described the task of creating a structured database of trial outcomes on the basis of data recorded in trial registries. Outcomes extracted from registries vary significantly in terms of their length, level of detail included in the definition of an outcome, and syntactic structure. The absence of uniformity in defining outcomes in registries makes the creation of a structured database a difficult task.

We described our first experiments on clustering of the extracted outcomes and the difficulties encountered. Due to the mentioned absence of uniformity in defining outcomes, finding an optimal number of clusters proved to be difficult in our current experiments.

We outlined some machine learning and rule-based methods that we consider useful for creating a database of outcomes. We propose to extract a complete definition of an outcome from the free-text descriptions in registries using a deep learning method, and to consequently extract information on time points, measurement methods etc. using simple rule-based techniques.

Acknowledgements. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 676207.

References

Beltagy I., Cohan A., Lo K. Scibert : Pretrained contextualized embeddings for scientific text. arXiv :arXiv :1903.10676 2019.

Blake, C., Lucic, A. Automatic endpoint detection to support the systematic review process. J. Biomed. Inform. 2015 ; 56, 42-56.

Clarke M., Williamson P. R. Core outcome sets and systematic reviews. Systematic Reviews 2016 ; 5 :11. doi :10.1186/s13643-016-0188-6.

Demner-Fushman D., Few B., Hauser S. E., Thoma G. Automatically identifying health outcome information in MEDLINE records. Journal of the American Medical Informatics Association 2006 ; 13(1):52-60.

Devlin J., Chang M., Lee K., Toutanova K. BERT : pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 2018 ; arXiv :1810.04805. URL http://arxiv.org/abs/1810.04805

Ferreira J. C., Patino C.M. Types of outcomes in clinical research, Jornal Brasileiro de Pneumologia 2017 ; 42-6 :5.

Koroleva A., Paroubek P. Automatic detection of inadequate claims in biomedical articles : first steps. Workshop on Curative power of Medical Data (MEDA) 2018. http://doi.org/10.5281/zenodo.1164680

Lee J., Yoon W., Kim S., Kim D., Kim S., So C. H., Kang J. Biobert : a pretrained biomedical language representation model for biomedical text mining, arXiv preprint arXiv :1901.08746 2019.

Lucic A., Blake C.L. Improving Endpoint Detection to Support Automated Systematic Reviews. AMIA Annu Symp Proc. 2016 ; 1900-1909.

Scheepers T., Kanoulas E., Gavves E. Improving Word Embedding Compositionality using Lexicographic Definitions. WWW 2018.

Summerscales R, Argamon S, Hupert J, Schwartz A. Identifying treatments, groups, and outcomes in medical abstracts. The Sixth Midwest Computational Linguistics Colloquium (MCLC 2009) 2009, Bloomington, IN, USA : Indiana University

Termier A., Rousset M.-C., Sebag M. Combining Statistics and Semantics for Word and Document Clustering. IJCAI'2001 Workshop on Ontology Learning 2001, Seattle, USA.

Résumé

Un essai clinique est une étude qui évalue les effets d'une ou de plusieurs interventions sur une population donnée en ce qui concerne certains « outcomes » - des variables contrôlées pour évaluer l'impact de l'intervention. Les outcomes sont l'une des caractéristiques essentielles d'un essai clinique. Les résultats sont définis par plusieurs aspects, tels que le nom de la variable contrôlée, l'outil de mesure utilisé, les points horaires, la métrique d'analyse, la méthode d'agrégation. Nous proposons de créer de manière semi-automatique une base de données structurée des outcomes des essais et des aspects les définissant, qui peut être utilisée comme support pour la tâche d'extraction automatique des outcomes ou pour le développement de « Core Outcome Sets » - ensembles de outcomes de base. Nous proposons d'utiliser les données des registres d'essais - des bases de données en ligne contenant des informations sur les essais cliniques, y compris les outcomes. Nous appliquons des techniques de traitement des langues supervisées et non supervisées pour décrire et analyser les outcomes extraits des registres.

Fouille de textes et repérage d'unités phraséologiques[1]

Paolo Frassi*, Silvia Calvi**, John Humbley***

*Université de Vérone
paolo.frassi@univr.it
**Université de Vérone
silvia.calvi@univr.it
***CLILLAC-ARP EA3967 Université de Paris
john.humbley@eila.univ-paris-diderot.fr

Résumé. Nous nous proposons de démontrer que les entités linguistiques de type *locution* et *collocation*, qui représentent deux types différents de phrasèmes, existent en langue de spécialité et qu'elles y sont particulièrement récurrentes. Pour ce faire, nous nous appuyons sur les données issues d'une extraction pilote à partir d'un sous-corpus concernant le domaine du marketing provenant d'un plus grand corpus en voie de constitution au Département des Langues de l'Université de Vérone. Les données quantitatives seront ensuite passées au peigne fin de l'analyse qualitative qui nous permettra d'identifier, sur la base des critères proposés par la Lexicologie Explicative et Combinatoire, des unités multiléxémiques de type *locution* et de type *collocation* ainsi que d'en proposer une annotation qui tienne compte de leurs propriétés syntactico-sémantiques en vue de leur représentation dans une base de données terminologique.

1. Introduction

Les humanités numériques sont au cœur des recherches menées auprès du Département des Langues et Littératures étrangères de l'Université de Vérone depuis l'obtention, en janvier 2018, d'un financement de la part du Ministère italien de l'éducation, de l'université et de la recherche pour le projet *Le Digital*

1 Paolo Frassi a rédigé les §§ 1, 2, 5, 6 ; Silvia Calvi a rédigé le § 4 ; John Humbley a rédigé le § 3.

Humanities applicate alle lingue e letterature straniere (*Les humanités numériques appliquées aux langues et littératures étrangères*).

L'une des équipes de ce projet (DIACOM, dont nous faisons partie) s'attache à mener une étude terminologique concernant le domaine du commerce international. En particulier, à la suite de la constitution d'un corpus de langue française concernant ce domaine spécifique, et à l'issue de l'extraction automatique de termes simples et complexes, nous nous proposons, sur le long terme, de constituer une base de données terminologique. Cette base contiendra des unités terminologiques sous forme de *réseau lexical* (sur le modèle du *Réseau Lexical du français*, cf. à ce propos Polguère 2014) et se propose ainsi de représenter non seulement les termes mais également les liens syntagmatiques et paradigmatiques que les termes nouent entre eux.

Un pareil objectif pose des problèmes d'ordre théorique et pratique quant à certaines unités terminologiques, notamment les phrasèmes, ou unités terminologiques complexes (Drouin 2002), que nous essayons de présenter dans cet article.

Après avoir brièvement présenté le corpus, nous nous proposons de parcourir la question de la modélisation d'unités phraséologiques en terminologie (Gouadec 1993 ; Temmerman 2000 ; Drouin 2002, L'Homme 2004 ; Elkin 2012). Nous montrons par la suite, à partir de quelques-uns des logiciels actuellement disponibles (notamment *TermoStat* et *TermSuite*), les possibilités d'extraction automatique d'unités phraséologiques. C'est à partir de la collecte de ces données que nous nous proposons de démontrer que les entités lexicales de type *collocation* ou *locution*, dans la définition proposée par Mel'čuk 2008 et 2013 ne sont pas l'apanage de la langue générale mais qu'elles existent également en langue de spécialité. Dans l'objectif plus spécifique de notre projet – la création d'une base de données de type *réseau lexical* – nous allons terminer par un modèle d'annotation qui permette de bien discrétiser ces deux types de phrasèmes dans l'objectif plus général d'en permettre une représentation adéquate.

2. Le corpus

Notre corpus se constitue, à l'heure actuelle, de 445 textes (environ 10 millions de tokens) recueillis à partir de trois critères : un critère chronologique, un critère thématique et un critère concernant la typologie textuelle.

Pour ce qui est du critère chronologique, l'objectif étant de prendre en compte l'évolution des unités terminologiques, les textes couvrent la période

qui va de 1850 à 2018. Cette diachronie a été répartie en trois synchronies successives : 1850-1914 ; 1945-1970 ; 1990-2018. Ces périodes ont été choisies car elles coïncident avec trois périodes-clés de l'histoire du commerce : 1) industrialisation (1850-1914) ; 2) boom économique (1945-1970) ; 3) nouvelles techniques de vente de biens et services, développement du marketing (1990-2018).

Quant au critère thématique, certains aspects, dont les textes se devaient de traiter, ont été pris en compte, et notamment : *macro-économie et économie internationale* ; le *secteur* ; le *type d'entreprise/activité d'entreprise*. Précisons que l'aspect *macroéconomie et économie internationale* inclut la politique commerciale et les relations internationales ; les aspects concernant la variation diatopique éventuelle ; les reflets sociaux du marketing ; l'aspect *secteur* inclut, d'une part, les produits et de l'autre les services. Quant aux *types d'entreprise/activité d'entreprise*, nous avons décidé de prendre en compte le management, le marketing, la logistique, le commerce électronique et le droit.

Et, finalement, sur la base du critère textuel, quatre types de textes ont été retenus : textes institutionnels ; textes scientifiques ou académiques ; textes relevant de la presse spécialisée ; textes provenant de la documentation d'entreprise.

C'est à partir d'une partie de ce corpus que nous avons mené l'étude pilote concernant le sous-corpus du marketing dont il sera question dans les §§ 4 et 5.

3. L'unité terminologique

3.1. Petite histoire des unités multilexémiques

La question de ce qui constitue l'unité terminologique se pose depuis que les linguistes, plutôt que les spécialistes de domaine, s'occupent de confectionner des dictionnaires spécialisés. On sait que les pionniers de la terminologie, Alfred Schlomann et Eugen Wüster (Lowe Schlomann et Wright 2006) pour n'en nommer que les principaux, étaient des ingénieurs et qu'ils s'entouraient d'experts du domaine concerné par chaque projet terminographique. Puisque leur approche était résolument onomasiologique, ils ne se posaient pas la question de savoir si une dénomination était composée d'un ou de plusieurs éléments. Ils partaient de l'unité conceptuelle et ils examinaient les différentes expressions langagières correspondantes. Mais à partir du moment

où le travail en terminographie a été confié à des langagiers, plutôt qu'à des experts, la perspective changeait : comment en effet savoir si un ensemble phraséologique renvoie à un concept pertinent du domaine étudié ? C'est ainsi, dans le cadre de la francophonie, que les responsables des principales banques de terminologie (La banque de terminologie du Québec, connue aujourd'hui sous le nom du *Grand Dictionnaire terminologique*, et *Termium Plus*, du niveau fédéral canadien) organisaient dès les années 1970 des colloques sur la définition de l'unité terminologique. Le critère onomasiologique était encore invoqué à cette époque, car les terminologues commençaient leur travail sur un micro-domaine en construisant « l'arbre du domaine », qui indiquait les concepts pertinents. En outre, ils envisageaient le travail de dépouillement en plusieurs étapes, qui comportait obligatoirement la consultation d'un expert, qui validait ou qui invalidait les termes retenus par le langagier. Cette division du travail est toujours d'actualité, pratiquée par exemple par les équipes travaillant sur des corpus importants. Plus généralement le terminologue recherche les traces linguistiques de (nouvelles) structurations conceptuelles, généralement dans des micro-domaines émergeants (cf. Blampain 1992). Il s'ensuit que la démarche implique la prise en compte d'une certaine latitude en ce qui concerne la définition de l'unité terminologique. Pour Estopà (2001) notamment, il est plus utile d'adopter une définition large et de parler d'*unité de signification spécialisée*, qui serait appréhendée différemment selon la catégorie d'utilisateur : c'est ainsi qu'un spécialiste du domaine, un traducteur, un rédacteur et un journaliste peuvent avoir des vues divergentes sur ce qui constitue un terme, surtout multilexémique. Selon ce point de vue, ce qui constitue un terme dépend de l'usage que l'on souhaite en faire.

Une étape de plus dans la réflexion sur l'unité terminologique fut franchie à partir du moment où l'automatisation de l'extraction est devenue une réalité. L'identification de l'unité terminologique est alors une priorité si l'on souhaite confier une partie du travail d'extraction (encore appelé dépouillement) à une machine : dans ce contexte une définition rigoureuse et objective de ce qui constitue un terme, surtout multilexémique, se pose avec acuité. Très tôt, les linguistes, généralement spécialistes du traitement automatique du langage, se sont rendu compte que les termes étaient très majoritairement des syntagmes nominaux. Dans le cadre d'une extraction terminologique portant sur anglais, par exemple, Daille *et al.* (1996 : 204) ont identifié 4 771 termes à deux éléments, (*multiple access, intermediate frequencies, modulating signal, co-polarized component, communication satellite*, *compression buffering* etc.), 1 045 à trois éléments et 168 à quatre éléments. Leur étude, portant sur

les termes composés de trois éléments, révèle les mécanismes en jeu pour la formation de ces syntagmes, à savoir l'insertion, la juxtaposition, la permutation ou/et la coordination. Les travaux sur la surcomposition syntagmatique sont particulièrement développés par Portelance (1987, 1998, 2000), qui identifie 48 matrices différentes. L'étude de ces termes « lourds » a été reprise et approfondie plus récemment sous le nom de *mégatermes* par Pecman (2018).

La description des termes multilexémiques ne serait pas complète si les relations sémantiques n'étaient pas prises en considération en même temps que leur description syntaxique. Il existe en réalité de nombreuses études réalisées depuis les années 1990 portant sur plusieurs langues (surtout sur l'anglais) dont nous ne mentionnons ici que les principales. Sager et Kageura (1994-1995) ont caractérisé les grandes catégories sémantiques des composantes de ces termes, tout comme Weissenhofer (1995), qui tentait de dégager les règles de combinaison qui rendent compte de cet aspect majeur de la création terminologique. Il postulait les catégories sémantiques suivantes : agent, causalité, composition, constitution, équation, instrument, méronymie, négation, origine, lieu, possession, propriété, but, quantité, ressemblance, temps. Au fil du temps, on a ajouté des catégories et on les sous-divisait encore, de telle sorte que Oster (2006), par exemple, postule plusieurs sous-catégories pour rendre compte de la localisation : la localisation à proprement parler : *Unterglasurmalerei (peinture sous verre)* ; la localisation relative : *Oberstempel (poinçon supérieur)* ; la localisation par rapport à l'origine : *Waidhaus-Feldspat (feldspath de Waidhaus)* (Oster 2006, 11).

3.2. Définir le domaine

Selon l'approche classique de la terminologie, le terme est défini non par sa forme, mais par le rôle qu'il joue dans un domaine de spécialité. Un terme n'est un terme que dans un domaine donné, et en changeant de domaine, on change de définition (*eau* en chimie n'a pas la même définition qu'en physique, par exemple). Ce point de vue a été relativisé par des linguistes, qui faisaient valoir que la polysémie existe aussi bien en langue de spécialité que dans la langue générale, et que la division en domaines étanches était non seulement contre-intuitive, mais handicapante, car elle empêche de rendre compte des concepts nomades, particulièrement important en termes d'innovation. Ces critiques étaient surtout formulées par les socio-terminologues, qui s'intéressaient avant tout aux conditions de la circulation des termes. Dans L'Homme (2004), en revanche, la prise en compte du domaine auquel appartient un terme est considérée comme une donnée.

3.3. Définition du domaine du commerce

Malgré un certain consensus pour la prise en compte des domaines en terminologie, il subsiste la difficulté de déterminer ce qui fait partie d'un domaine donné, et ce qui n'en fait pas partie. Une solution souvent adoptée est de partir des catégories des systèmes de classification en usage en documentation, en particulier de la Classification décimale universelle (CDU). Wüster s'en est servi pour classifier l'ensemble des termes de son *Dictionnaire de la machine-outil* (Wüster 1968).

En ce qui concerne le domaine du commerce il est possible d'adopter une définition large ou étroite. Certaines branches du commerce se sont plus développées que d'autres en fonction de leur importance dans l'organisation de ce secteur clé de l'économie. Le marketing, de nos jours le marketing électronique, revêt à ce titre une importance particulière. C'est au terminologue de rendre compte du matériau linguistique dont cette manifestation est faite, de rendre compte de la constitution d'un nouveau réseau conceptuel exprimé par les termes émergents, et de compléter ainsi les dictionnaires spécialisés existants. Ceci se fait en s'appuyant sur les manifestations réelles des discours spécialisés, c'est-à-dire entre spécialistes, et d'interface, entre spécialistes et publics plus larges.

3.4. Tri par le choix du corpus

De nombreux terminologues (cf. Cabré 2007-2008) insistent sur l'importance de bien calibrer le corpus qui sera exploité pour l'extraction : celui-ci doit être construit de telle façon qu'il reflète les dimensions pragmatiques de la variation terminologique. En effet, on privilégie les textes de recherche, car il s'agit d'experts qui s'adressent à d'autres experts, garantissant ainsi la fiabilité des contextes définitoires. L'inclusion de textes d'initiation, comme les manuels scolaires ou de premier cycle universitaire, est recommandée moins pour l'inclusion de terminologie récente, mais plutôt pour assurer la présence d'une terminologie consensuelle : le critère pragmatique étant la communication entre experts et futurs experts. Les textes de vulgarisation constituent une troisième tranche de corpus, et correspondent à la communication spécialistes-non spécialistes.

3.5. Les ressources terminologiques dans le domaine du commerce

Les ressources en terminologie commerciale en langue française comptent deux monuments : Dancette et Réthoré (2000) et Van Dyck *et al* (2001). Ce dernier est surtout un dictionnaire de décodage, tandis que celui de la vente au détail est prévu pour l'encodage en plus. La migration sur la Toile est désormais assurée, comme le démontre le dictionnaire de Dancette (2007) sur la mondialisation du travail ainsi que l'article panorama de Temmerman (2003).

Si l'on s'intéresse plus précisément au marketing on constate que de nombreux dictionnaires spécialisés existent. Le SUDOC, catalogue collectif des bibliothèques universitaires françaises, fait état de près de vingt dictionnaires papier du français du marketing. Il s'agit essentiellement de dictionnaires pédagogiques, à la nomenclature limitée généralement à environ 300 notions. Celui qui se trouve dans le plus grand nombre de bibliothèques (102 en mai 2019), est l'ouvrage d'Albertini *et al*. ([2001] 2008).

Les ressources électroniques sont bien moins aisées à saisir. Il existe une multitude de « glossaires » du marketing sur la Toile dont la qualité est variable, mais parfois l'œuvre de spécialistes. C'est le cas de *Définitions marketing*, ressource, conçue et réalisé par Bertrand Bathelot (2019), professeur agrégé de marketing, formateur indépendant et surtout spécialiste du domaine. Il est de taille considérable car il comporte plus de 6000 articles (soit vingt fois plus que les dictionnaires papier des bibliothèques, eux-mêmes composés de définitions, exemples et illustrations sous forme de vidéos, diaporamas et autres images et figures), ainsi qu'un réseau d'hyperliens permettant une circulation à l'intérieur du dictionnaire. Celui-ci est par ailleurs organisé par regroupements thématiques, appelés *Glossaires*. Par exemple, le marketing sportif regroupe 40 articles. Les termes sont à plus de 80 % de nature multilexémique : cf. *zone de chalandise, marketing sportif, point de vente*. Les définitions en revanche sont peu formalisées et l'analyse linguistique ne fait pas partie des objectifs de cet outil, comme le montre celle de *hashtag*, illustrée ci-dessous :

HASHTAG : Un hashtag est mot ou groupe de mot suivant le caractère # dans un tweet. Créé à l'initiative du concepteur du message un hashtag est cliquable et permet au lecteur d'être redirigé vers des tweets traitant du même sujet. Dans le tweet : « je viens d'arriver à #SaintTropez », un clic sur le hashtag SaintTropez permet d'accéder à d'autres tweets sur le même sujet. Lorsqu'on utilise Twitter à des fins de promotion, il est conseillé d'utiliser les hashtags pour bénéficier d'une audience recherchant des tweets à partir de mots clés.

On note que l'auteur de la définition précise bien dans l'exemple la pertinence de ce terme dans le contexte du commerce électronique.

Nous avons vu qu'il existe de très nombreuses tentatives de catégoriser l'unité terminologique multilexémique, étape indispensable de l'extraction et de la description de termes destinés à constituer une base de données terminologiques du français du marketing. C'est le but des lignes qui viennent.

4. L'extraction d'unités terminologiques : étude pilote

Dans cette section nous présentons les résultats d'une étude pilote qui nous a conduit à une première extraction d'unités phraséologiques. Nous n'avons considéré qu'une seule partie du corpus : des textes allant de 1990 à 2018 qui appartiennent au sous-domaine du *type d'entreprise/activité d'entreprise*, notamment la partie des textes concernant le *marketing*. Il s'agit de 18 textes (soit 776 348 *word tokens*) de nature différente : textes scientifiques et académiques (50 %), articles de la presse spécialisée (33 %) et textes institutionnels (17 %).

Nous avons commencé par la conversion des *pdf* en *txt* pour pouvoir insérer les textes dans les extracteurs automatiques. Nous avons aussi décidé de supprimer de manière manuelle toutes les parties qui n'avaient aucun intérêt linguistique, comme par exemple les frontispices, les index, les titres, les notes en pied de page, les grilles, les bibliographies etc. La suppression de ces parties visait à réduire le bruit dans les résultats à analyser.

Ensuite, pour ce qui est des extracteurs notre choix est tombé sur les deux logiciels suivants : *TermoStat*, conçu et réalisé au sein de l'Observatoire de Linguistique Sens-Texte de l'Université de Montréal (Drouin 2003) et *TermSuite*, développé à l'Université de Nantes (Cram et Daille 2016). L'emploi de deux extracteurs nous a permis de comparer les résultats et, par conséquent, d'augmenter leur fiabilité. À partir du logiciel *TermoStat* nous n'avons extrait que les unités multilexémiques dont le premier élément était un substantif. Nous avons ensuite mené le même type d'extraction avec *TermSuite*.

Pour résumer, comparer et confirmer les résultats, nous avons créé une grille Excel dans laquelle nous avons repris les informations suivantes : 1. les candidats de regroupement, c'est-à-dire toutes les unités multilexémiques dont le premier élément était un substantif ayant pour *TermoStat* une fréquence d'au moins 10 occurrences ; 2. la matrice ; 3. la fréquence dans les deux logiciels. Nous avons inclus dans notre grille 517 candidats de regroupement que nous avons ensuite passés en revue de manière manuelle afin de retenir uni-

quement les unités phraséologiques appartenant au domaine du marketing. Nous avons ainsi exclu un nombre important de termes (274), comme par exemple *revue de littérature*, *cadre théorique*, *événement sportif*. Nous avons, ensuite, éliminé toutes les unités multilexémiques (63) qui n'étaient pas des unités phraséologiques, comme par exemple *relation vendeur-client, marque choisie, entreprise tunisienne*.

À la suite de ce double filtrage manuel nous avons obtenu 180 unités phraséologiques du domaine du marketing. Ce n'est que pour ces unités phraséologiques que nous avons mené une étude de nature qualitative qui nous a permis de classer ces unités multilexémiques dans les typologies majeures d'unités phraséologiques - locutions et collocations.

5. Les locutions et les collocations en langue de spécialité

5.1. Modèle théorique

Notre modélisation des unités phraséologiques ou phrasèmes (locutions et collocations) s'appuie sur le classement des unités multilexémiques non libres qu'en propose Mel'čuk 2008 et 2013 dans le cadre de la Lexicologie Explicative et Combinatoire (dorénavant LEC). En particulier ce classement se base sur deux paramètres de base des unités multilexémiques, qui permettent déjà de distinguer les unités multilexémiques libres des unités multilexémiques non libres :

1. leur liberté sur l'axe paradigmatique
2. leur liberté sur l'axe syntagmatique

Ces deux critères s'appliquent de manière différente aux diverses catégories d'unités multilexémiques non libres.

En général, toutes les unités multilexémiques non libres sont contraintes sur l'axe paradigmatique, car il est impossible de remplacer l'une des composantes du phrasème par une unité lexicale ou une expression suffisamment synonymique (voir exemples, § 5.2).

En revanche, il existe une différence sur l'axe syntagmatique – à savoir l'application de règles suffisamment générales pour l'agencement des différentes composantes de chaque unité phraséologique. Ainsi, pour les locutions, le locuteur n'a aucune liberté – ni aucun pouvoir – sur l'axe syntagmatique : une locution est donnée d'avance par le système langue comme un tout pré-confectionné. Pour les collocations, qui se composent de deux lexies

différentes (cf. § 5.2), il existe un certain degré de liberté dans l'application de règles grammaticales suffisamment générales (ex. : *remède efficace*, *remède très efficace*).

5.2. Locutions et collocations

Ces mêmes études de Mel'čuk distinguent trois types de locutions et deux types de collocations. Les trois types de locutions se situent sur une échelle allant du +/- opaque au +/- transparent.

Sont complètement opaques les locutions fortes, dont le sens global n'est pas compositionnel car les deux composants sont sémantiquement opaques (ex. : *casser sa pipe*). Sont complètement transparentes les locutions faibles (ou quasi-compositionnelles) dont les deux composants sont sémantiquement transparents (ex. : *centre commercial*). En dépit de leur transparence sémantique, ces unités multilexémiques sont bel et bien des locutions : pour preuve, elles sont assujetties aux deux contraintes qui jouent sur l'axe syntagmatique et paradigmatique ; en effet, il est impossible de remplacer l'un des deux constituants par un autre suffisamment synonymique (ex. **centre de commerce*, **cœur commercial, *foyer commercial*) ; en outre, cette lexie est prise par le locuteur comme un seul et unique bloc, sur lequel il ne peut intervenir librement du point de vue grammatical (ex. : **centre assez/très/peu commercial*).

Les locutions semi-compositionnelles constituent, quant à elles, un degré moyen d'opacité/transparence : dans ce type de locutions un constituant est transparent alors que l'autre est opaque (ex. : *fruit de mer*).

Quant à la collocation, rappelons que cette entité phraséologique se compose de deux éléments distincts, une base et un collocatif ayant des statuts sémantico-sytaxiques différents car si, d'une part, la base est toujours autonome (elle peut fonctionner dans la langue indépendamment de son collocatif) le collocatif ne peut fonctionner, dans la plupart des cas, sans sa base : par exemple dans *pleuvoir des cordes*, *pleuvoir* est autonome, alors que *des cordes* s'active dans le sens de 'très, intense' uniquement s'il est associé à *pleuvoir*. C'est pour cette raison que l'on dit que, dans une collocation, la base contrôle son collocatif (cf. à ce propos : Hausmann 1989 ; Hausmann et Blumenthal 2006 ; Polguère et Mel'čuk 2006).

La LEC prévoit deux types de collocations, suivant leur formalisation via le système des fonctions lexicales : d'une part, les collocations formalisées par des fonctions lexicales standard, de l'autre les collocations formalisées par des

fonctions lexicales non standard. Nous ne nous attardons pas ici sur la notion de *fonction lexicale* ; nous nous limitons à rappeler que celle-ci est la transposition, sur le plan lexical, d'une fonction mathématique (*f(x) = y*) : à la suite de cette transposition, les deux variables de la fonction (l'argument et la valeur) correspondent à deux lexies d'une langue donnée et la fonction à proprement parler représente un lien de type syntagmatique ou paradigmatique particulièrement récurrent et généralisable dans la plupart des langues du monde. Pour une vue plus détaillée des fonctions lexicales voir Mel'čuk, Clas et Polguère 1995 ; Wanner 1996.

Dans le cas des collocations formalisables par une fonction lexicale standard, un collocatif s'associe à une base à partir d'un lien syntagmatique assez récurrent dans une langue donnée ainsi que dans la plupart des langues du monde. Par exemple, dans de nombreuses langues, nous avons des exemples de collocations dans lesquelles un collocatif s'associe à une base pour signifier 'très', 'intense' : *faire noir comme dans un four* (fr.) ; *pleuvoir des cordes* (fr.) ; *it's raining cats and dogs* (ang.).

Il existe, en revanche, des collocations dans lesquelles la base ne s'associe pas à un collocatif pour exprimer un sens très récurrent (ex. : *café noir*) : il s'agit de collocations représentées par les fonctions lexicales non standard.

La limite très faible entre ce second type de collocations et les locutions semi-figées a déjà été relevée Mel'čuk, Clas, Polguère (1995) :

> il existe en effet un nombre imprévisible de locutions semi-figées (=collocations), qui d'une part, sont strictement du même type que les locutions « lexico-fonctionnelles », mais qui, d'autre part, ne peuvent pas être décrites par les FL standard- puisque leur sens est trop spécifique et donc non généralisable (Mel'čuk, Clas, Polguère 1995, 150)

Ce type d'unités phraséologique est particulièrement productif dans les langues de spécialité (cf. Mel'čuk, Clas, Polguère 1995, 151). Nous croyons que, pour plusieurs raisons liées à l'application de restrictions syntactico-sémantiques particulières, ces unités phraséologiques s'apparentent davantage aux locutions faibles qu'aux collocations.

Par exemple, dans de nombreux cas, dans les collocations une base peut s'associer à plusieurs collocatifs pour exprimer un sens donné : ex. : *faire noir comme dans un four/comme dans la gueule d'un loup/comme dans un tunnel*. Les collocations dites *non standard* présentent très difficilement cette possibilité (*café noir, *café sombre*) – propriété qu'elles partagent avec les locutions.

En outre, dans les collocations, du fait de leur relative liberté grammaticale, il est impossible d'insérer du matériau linguistique : par exemple on peut dire *un remède efficace* ou *un remède très efficace*. Si on reprend une collocation dite *non standard*, nous remarquons qu'elle ne possède pas cette liberté sur le plan syntagmatique : *un café noir/*un café très noir.*

5.3. Phrasèmes et langues de spécialité

Les langues de spécialité présentent des types de termes complexes correspondant aux deux entités phraséologiques qui se présentent dans la langue générale que nous venons de présenter au § 5.2.

L'extraction que nous avons menée à travers notre étude pilote, nous a donné des locutions fortes (ex. *marge arrière*), des locutions semi-compositionnelles (ex. : *marketing viral*) et des locutions quasi-compositionnelles (ex. : *valeur ajoutée*). Ces unités multilexémiques sont assujetties aux mêmes contraintes sémantico-syntaxiques que les locutions dans la langue générale.

L'extraction a mis en relief également des unités multilexémiques de type *collocation*, avec les mêmes sémantismes que l'on peut retrouver dans la langue générale : par exemple, le sens 'bon' (*client fidèle*, *attitude positive*), le sens 'tel qu'il faut' (*performance objective*, *effet attendu*) ou le sens 'très, intense' (*marque forte*, *valeurs extrêmes*).

Remarquons que les collocations ne font pas la majorité des entités phraséologiques : les locutions se taillent la part du lion avec 90 % d'occurrences.

Nous ne croyons pas, pour autant, que les collocations se présentent avec une fréquence faible dans les langues de spécialité ; la raison de cette pénurie dépend plutôt du fait que les extracteurs actuellement disponibles se concentrent sur des termes complexes de nature nominale (souvent à base nominale et à collocatif adjectival) et négligent, dans la plupart des cas, les termes complexes de type *collocation* à collocatif verbal (ex. : *exercer un commerce* ou *gérer un commerce*).

Une remarque s'impose aussi à propos des locutions : la plus grande partie sont des locutions quasi-compositionnelles (82 %) ; suivent les locutions semi-compositionnelles (15 %) et les locutions fortes (3 %). L'explication d'un pourcentage aussi élevé des locutions quasi-compositionnelle se retrouve dans deux raisons différentes. D'une part, nous avons décidé de faire rentrer les collocations de type *non standard* dans l'ensemble des locutions faibles car les collocations non standard partagent les mêmes propriétés syntacti-co-sémantique que les locutions faibles. Cela n'explique que partiellement le

nombre important de locutions faibles. Une seconde raison à la base de la prolifération des locutions faibles se retrouve dans l'activité de lexicalisation propre aux langues de spécialité : celles-ci se proposent souvent de dénommer des nouveaux *realia* par des unités lexicales qu'elles forgent *ad hoc*. Ces unités lexicales sont le plus souvent des locutions faibles car elles ont : 1) un haut degré de cohésion, surtout au niveau syntaxique, et sont immédiatement perçues par le locuteur comme un seul bloc sémantique ; 2) un haut degré de transparence sémantique, qui permet au locuteur d'accéder plus facilement au sens de l'unité multilexémique.

5.4. Quelle annotation ?

Notre réflexion d'ordre linguistique sur les unités phraséologiques et sur leur catégorisation a des débouchées d'ordre terminographique : nous comptons, en effet, représenter de manière adéquate, et dans le respect de leurs propriétés syntactico-sémantiques réciproques, les types divers d'entités phraséologiques dans la base de données terminologique qui va être constituée à la suite de l'extraction automatique des termes. En amont de notre représentation qui, comme nous l'avons précisé dans l'introduction, va suivre le modèle des bases de données de type *réseau lexical*, nous avons prévu une annotation qui puisse tenir compte de deux aspects :

1. La nature et les propriétés de chaque entité phraséologique (locution et collocation) et de chacune des sous-catégories (locutions fortes, locutions semi-compositionnelles, locutions quasi-compositionnelles).
2. La possibilité de créer des liens paradigmatiques et syntagmatiques entre les entités lexicales.

Une annotation permettant une représentation des connaissances sous forme structurée en RDF (*Resource Description Framework*) pourrait conjuguer les deux aspects : ce type d'annotation permet, en effet, d'annoter chaque type d'entité lexicale et de discrétiser ainsi une locution d'une collocation ou, encore, une locution forte d'un locution semi-compositionnelle ou, finalement, une collocation exprimant le sens 'très, intense' d'une collocation exprimant le sens 'bon' ; il permet, en outre, de créer des liens de type paradigmatique ou de type syntagmatique.

L'exemple suivant, qui est une ébauche que nous proposons sur la base du fonctionnement de l'annotation en RDF, montre sans contexte la possibilité de distinguer clairement non seulement les types divers d'entités phraséologiques mais, également, des sous-catégories éventuelles au sein de chaque

type d'entités. La première série concerne trois locutions, dont une locution forte (*marge arrière*), une locution semi-compositionnelle (*marketing viral*) et une locution faible (*orientation client*). Puisque, sur la base des propriétés présentées plus haut, la locution est considérée comme une seule lexie, le type de locution (forte, semi-compositionnelle, faible) est suffisante pour identifier l'entité phraséologique :

ID …
label : marge arrière
type : locution forte
ID…
label : marketing viral
type : locution semi_compositionnelle
ID …
label : orientation client
type : locution faible

Pour ce qui est des collocations, en revanche, la précision de ses deux composantes (la base et le collocatif) est de mise dans l'annotation ; ce type d'annotation permet non seulement de discrétiser une locution d'une collocation au niveau du *type* d'entité lexicale, mais de respecter l'identité de chaque entité phraséologique – dans ce cas la collocation – par l'expression de ses deux composantes, la base et le collocatif. Elle permet, en outre, d'identifier le type de collocation sur la base du sémantisme qu'elle véhicule (dans le cas qui suit, le sémantisme 'très, intense' est codé par le nom de la fonction lexicale Magn, l'une des fonctions lexicales syntagmatiques identifiées par la LEC) :

ID …
label : marque forte
type : collocation Magn
base ID … : marque
collocate ID … : forte

6. Conclusions

Nous avons essayé de démontrer, à travers l'application de critères proposés par la LEC, que des entités lexicales de type *locution* ou *collocation* se manifestent avec une fréquence importante en langue de spécialité et qu'elles ont droit de cité dans des bases de données terminologiques. Ces mêmes entités phraséologiques ont souvent été négligées par les études terminologiques

et par les bases de données terminologiques, qui privilégient la distinction entre termes simples et termes complexes

L'emploi d'un annotateur de type RDF, en outre, nous a permis de proposer un modèle d'annotation respectant, pour chacune des entités phraséologiques prises en compte, les propriétés qui leurs sont propres ; cette annotation est en mesure de permettre, par conséquent, une représentation adéquate de chacune de ces types de données phraséologiques, ainsi que de créer des liens paradigmatiques et syntagmatiques avec d'autres types de termes – simples ou complexes. Non seulement : l'annotation permettra également d'avoir accès à un ensemble de données quantitatives qui vont pouvoir intéresser le linguiste. Par exemple, nous comptons introduire des types de requêtes concernant l'ensemble des locutions fortes, ou l'ensemble des locutions quasi-compositionnelles. Ou, également, des requêtes plus complexes, comme par exemple le pourcentage des collocations véhiculant le sens 'très, intense' par rapport à la totalité des collocations ou, plus généralement, par rapport à la totalité des entités phraséologiques. Les linguistes sont parfaitement conscients que les données quantitatives ne sont jamais suffisantes pour interpréter des phénomènes linguistiques ; il n'en reste pas moins qu'elle s'avèrent particulièrement utiles pour photographier le comportement de certains types d'entités – comme les entités phraséologiques – pour pouvoir effectuer une comparaison entre la langue générale et la langue de spécialité ou, encore, entre plusieurs domaines ou sous-domaines touchés par le travail terminologique.

Références

Albertini *et al*. [2001]2008. *Dictionnaire du marketing*. Paris : Vuibert.

Bathelot, Bertrand. 2019. *Définitions marketing*. Dernier accès : mai 2019. https://www.definitions-marketing.com/

Blampain, Daniel. 1992. « Traduction et écosystèmes terminologiques ». *Terminologie et traduction* 2/3 : 457-466.

Cabré, Maria Teresa. 2007-2008. « Constituer un corpus de textes de spécialité ». *Cahiers du CIEL*. Dernier accès : mai 2019. www.eila.univ-paris-diderot.fr/_media/recherche/clillac/ciel/cahiers/…/04-cabre.pdf

Cram, Damien ; Daille, Béatrice. 2016. « Terminology extraction with term variant detection ». In *Proceedings of ACL-2016 System Demonstration*, 13-18. Dernier accès : mai 2019. https://www.aclweb.org/anthology/P16-4003

Daille, Béatrice ; Habert, Benoît ; Jacquemin, Christian ; Royauté, Jean. 1996. « Empirical observation of term variations and principles for their description ». *Terminology* 3/2 : 197-257.

Dancette, Jeanne. 2007. « La mondialisation du travail : des pratiques sociales à la terminologie et de la terminologie à l'usage ». *Revue internationale sur le travail et la société* 5/2 : 64-83.

Dancette, Jeanne ; Réthoré, Christophe. 2000. *Dictionnaire analytique de la distribution*. Montréal : Les Presses universitaires de Montréal.

Drouin, Patrick. 2002. « Acquisition automatique des termes : l'utilisation de pivots lexicaux spécialisés ». Thèse de doctorat, Université de Montréal. Dernier accès : mai 2019. http://olst.ling.umontreal.ca/pdf/DrouinPhD2002.pdf.

Drouin, Patrick. 2003. « Term extraction using non-technical corpora as a point of leverage ». *Terminology* 9/1 : 99-117.

Elkin, Peter L. (sous la direction de). 2012. *Terminology and terminological systems*. London : Springer Science & Business Media.

Estopà Bagot, Rosa. 2001. « Les unités de signification spécialisées élargissant l'objet du travail en terminologie ». *Terminology* 7/2 : 217-237.

Gouadec, Daniel. 1993. *Terminologie et phraséologie : acteurs et aménageurs.* Paris : La Maison du Dictionnaire.

Hausmann, Franz Josef. 1989. « Le dictionnaire des collocations ». In *Wörtherbücher. Ein internationales Handbuch zur Lexicographie*, sous la direction de Franz Josef Hausmann *et al.*, 1010-1019. Berlin/New York : Walter de Gruyter, vol. 1.

Hausmann, Franz Josef ; Blumenthal, Peter. 2006. « Présentation : collocations, corpus, dictionnaires ». *Langue française* 150 : 3-13.

L'Homme, Marie-Claude. 2004. *La terminologie : principes et techniques*. Montréal : Les Presses universitaires de Montréal.

Lowe Schlomann, Elizabeth ; Wright, Sue Ellen. 2006. « The Life and Works of Alfred Schlomann : Terminology Theory and Globalization ». In *Modern approaches to terminological theories and applications*, sous la direction de Picht Heribert, 153-162. Berne : Peter Lang, collection Linguistic insights.

Mel'čuk, Igor ; Clas, André ; Polguère, Alain. 1995. *Introduction à la lexicologie explicative et combinatoire*. Paris : Duculot.

Mel'čuk, Igor. 2008. « Phraséologie dans la langue et dans le dictionnaire ». *Repères & Applications* VI : 187-200.

Mel'čuk, Igor. 2013. « Tout ce que nous voulions savoir sur les phrasèmes mais… ». *Cahiers de lexicologie* 102/1 : 129-149.

Oster, Ulrike. 2006. « Classifying domain-specific intraterm relations : A schema-based approach ». *Terminology* 12/1 : 1-17.

Pecman, Mojca. 2018. *Langue et construction des connaisSENSes : énergie lexico-discursive et potentiel sémiotique des sciences*. Préface de Marie-Claude L'Homme. Paris : L'Harmattan.

Polguère, Alain ; Mel'čuk, Igor. 2006. « Dérivations sémantiques et collocations dans le DiCo/LAF ». *Langue française* 150 : 66-83.

Polguère, Alain. 2014. « From Writing Dictionaries to Weaving Lexical Networks ». *International Journal of Lexicography* 27/4 : 396-418.

Portelance, Christine. 1987. « Fertilisation terminologique ou insémination terminologique artificielle ? ». *Meta* 32/3 : 356-360.

Portelance, Christine. 1998. « Figement lexical et flexibilité paradigmatique des vocabulaires specialisés ». In *Le figement lexical* sous la direction de Salah Mejri *et al*, 259-270. Tunis : Ceres.

Portelance, Christine. 2000. « Le statut exceptionnel de l'adjectif dans le syntagme dénominatif ». In *La Traduction : diversité linguistique et pratiques courantes. Actes du Colloque international 'Traduction humaine, Traduction automatique, Interprétation'* sous la direction de Salah Mejri *et al.*, 149-158. Tunis : Orbis édition, "Série linguistique" 11.

Sager, Juan Carlos ; Kageura, Kyo. 1994/1995. « Concept classes and Conceptual Structures : their Role and Necessity in Terminology ». *ALFA, Actes de langue française et de linguistique* 7 : 191-216.

Temmerman, Rita. 2000. *Towards new ways of terminology description : the socio-cognitive approach*. Amsterdam/Philadelphia : John Benjamins Publishing.

Temmerman, Rita. 2003. « Innovative methods in specialised lexicography ». *Terminology* 9/1 : 117-135.

Van Dyck, Jan ; Binon, Jean ; Verlinde, Serge ; Bertels, Ann. 2001. *Dictionnaire d'apprentissage du français des affaires*. Paris : Didier.

Wanner, Leo (éd.). 1996. *Lexical Functions in Lexicography and Natural Language Processing*. Amsterdam/Philadelphia : John Benjamins Publishing.

Weissenhofer, Peter. 1995. *Conceptology in Terminology Theory, Semantics and Word-formation*. Vienne : TermNet, "IITF SERIES" 6.

Wüster, Eugen. 1968. *Dictionnaire multilingue de la machine-outil*. Londres : Technical Press.

Abstract

The aim of this paper is to show that those linguistic entities known as idioms and collocations, which represent two different kinds of phrasemes, exist in language for special purposes, where they both occur frequently. In order to support this claim, data were extracted from a pilot subcorpus made up of marketing texts drawn from a larger corpus, which is in the process of being compiled at the Language Department of the University of Verona. The quantitative data were then analysed from a qualitative point of view, using criteria established in Explanatory and Combinatorial Lexicology to identify multilexemic units as either idiom or collocation types and to put forward a notation which accounts for their syntactico-semantic properties with view to their presentation in the termbase.

Dealing with specialised co-text in text mining : Verbal terminological collocations

Margarida Ramos* **, Rute Costa*, Christophe Roche** *

* NOVA CLUNL, Centro de Linguística da Universidade NOVA de Lisboa Avenida de Berna 26-C, 1069-061 Lisboa – Portugal
** Condillac Group – Listic Lab. Université Savoie Mont-Blanc Campus Scientifique, 73 376 Le Bourget du Lac cedex – France
mvramos@fcsh.unl.pt, rute.costa@fcsh.unl.pt, christophe.roche@univ-savoie.fr

Abstract. The aim of this paper is to organise lexical and conceptual knowledge by analysing a domain-specific corpus. The domain we focus on is the cork industry. Through the analysis of the corpus, we have found that some common verbs in Portuguese, such as "choose" and "separate" acquire a specialised value in the field under study. This was the starting point for the analysis of the terminological collocations where verbs are the core constituents. For the analysis of these verbal terminological collocations, we used natural language processing techniques, building simple and complex CQL structures with RegEx. The outcome of this analysis permits us to introduce a distinction between polylexical terms and terminological collocations. The terminological collocation is a reality of great relevance in specialised discourse, but unlike terms, it is not defined by conceptual criteria, but by morphological and syntactic criteria.

1. Introduction

The aim of this paper is to organise lexical and conceptual knowledge by analysing a domain corpus. The domain we shall focus on is the cork industry. Our work develops around a number of activities that are divided into 3 subsectors based on cork-related tasks : (1) the preparation of cork ; (2) its transformation ; and (3) the agglomeration of cork products. The texts that make up the corpus report on these activities reflecting the actions required to carry out the specific tasks on each of the sub-sectors identified above. By analysing the corpus, we have found that the verbs "escolher" [choose] and "separar" [separate] acquire a specific specialised value in the field under study and will

therefore be the starting point for the analysis of the terminological collocations where verbs are the core constituents.

In order to analyse these verbal terminological collocations (VTC), we have used natural language processing techniques related to text analysis envisaged as one of the tasks involved in text mining :

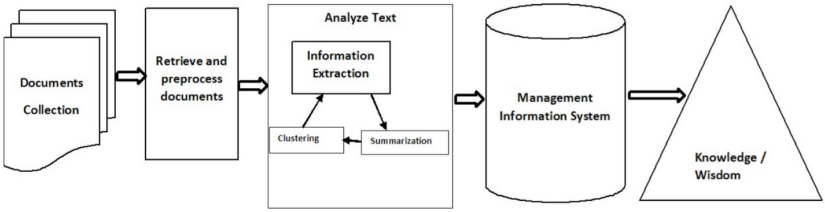


Fig. 1 - *Ramzan, Talib, Muhammad, K. Hanify, Shaeela, Ayeshaz and Fakeeha, Fatima. 2016.*

The texts under analysis are part of a corpus on the topic of cork designed for this purpose. We have used Sketch Engine[1] for text analysis aiming at observing verbal terminological collocations. Our work is based on the previously identified verbs as well as the co-texts in which the verbs occur. By co-text we mean the linguistic sequences that are next to the verbs that were identified for analysis. Using CQL[2], where RegEx (Regular Expressions) are used, we start from the morphosyntactic pattern [[Vtr] + [N]]TVC, where [V] is either the verb "escolher" [choose] or "separar" [separate], and [N] is a term[3] whose morphosyntactic structure is simple or complex. In Table 1, we present 8 examples found in the corpus corresponding to [N], which is a term – monolexical or polylexical – in the pattern of the structure under analysis :

	Type	Example - PT	EN [literal translation]
1	N	rolha	stopper
2	N+N	cortiça secundeira	secondary cork
3	N+Adj	rolha chanfrada	chamfered stopper
4	N+Prep+N	rolha de cortiça	cork stopper

1 https ://www.sketchengine.eu/
2 Corpus Query Language.
3 a *designation* that represents a *general concept* by linguistic means (ISO 1087 :2019)

5	N+N+Adj	cortiça virgem planificada	planned virgin cork
6	N+Adj+Adj	rolha preparada seca	dry prepared stopper
7	N+Adj+Prep+N	rolhas naturais com qualidade	natural stoppers with quality
8	N+Prep+N+Adj	rolhas para utilizações industriais	stoppers for industrial usage

Tab. 1 – *Terminogenic matrix.*

In this paper, we will make some considerations on: (i) verbal terminological collocations; (ii) domain specificities; and (iii) automatic corpus processing using Sketch Engine. Data will be extracted using CQL[4], where RegEx (Regular Expressions) are used, based on a previously prepared terminogenic matrix. The results obtained will be analysed according to the double dimension of Terminology (Costa, 2013).

2. Description of the activities involved in the cork domain

The corpus under analysis is made up of texts produced within the cork industry. As stated in the Introduction, the activity of this industry is divided into 3 subsectors, which include the activities of preparation, transformation, and agglomeration of cork products. The activity of the preparation of cork has to do with slicing [traçamento], stacking [empilhamento], boiling [cozedura], and stabilising [estabelização] the cork. The transformation of cork corresponds to activities that are associated to the manufacture of cork stoppers, which include the production and finishing [acabamento] of cylindrical batons of granulated cork for the manufacture of agglomerated cork stoppers and component-parts of technical stoppers, as well as natural and technical stoppers. The activities of the agglomeration of cork products include the production of materials for the construction industry and for the automobile and aeronautical sectors, among others.

We will focus on the subsector of transformation since the object under study is the production of natural cork stoppers. From the extraction of cork to the final product, several stages are needed, depending on the type of stopper one wants to produce. The overall process of the production of the cork stopper is divided into 3 stages, namely debarking [descortiçamento], manufacturing the stopper [fabrico da rolha], and finishing the stopper [acabamento

4 Corpus Query Language.

da rolha], where each stage encompasses different processes as illustrated below :

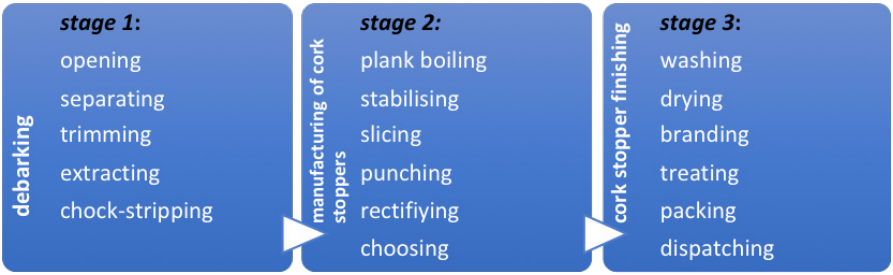


Fig. 2 – *Production of cork stoppers and its different stages (Nunes, 2013).*

The stages of debarking and finishing the stopper are the same for natural and agglomerated cork stoppers. The manufacturing of cork stoppers [fabrico da rolha] shown in Figure 2 relates exclusively to natural cork stoppers, while agglomerated cork stoppers undergo other processes that are included in the agglomeration activity, a process that will not be treated in this context.

The production of natural cork stoppers is one of the transformation activities. This type of stopper is obtained from a thick rectangular-shaped piece of cork named "stripe" [rabanada] through "punching" [brocagem[5]]. Experts explain that stripes are punched [brocadas], and to get those stripes [rabanadas] into a rectangular shape, the planks had to be previously sliced [rabaneadas]. At this stage, after being punched from the stripe, the stopper is only a semi-manufactured product, quite far from being a finished product.

A semi-manufactured natural cork stopper undergoes additional operations until it is a finished product. This is where the finishing process plays its role in the transformation activity. Cork stoppers may be sold with a semi-finished or finished status. The client acquires them (a winery, for instance) either unready or ready to be used, depending on the client's purposes or means to finish the stoppers. Briefly, a semi-finished stopper is a stopper that was submitted to any finishing treatment [tratamento de acabamento] of the finishing process [processo de acabamento], such as rectifying [rectificação], washing [lavação], and subsequent drying [secagem], except for the final treatment

5 Punching is the term used for the manual, semi-automatic or automatic process of perforating the strips of cork with a drill (APCOR : https ://www.apcor.pt/en/cork/processing/industrial-path/natural-cork-stoppers/).

[tratamento final]. At this point, the unready-for-use-stopper is either sold, packed and transported, or continues through the finishing process, until it is ready to be used. To be considered a finished product, the stopper must undergo the final treatments, which are branding [marcação] and/or surface coating treatment.

3. Domain-specific corpus : creation

To attain our terminological goals, we decided to build a domain-specific corpus, i.e., a corpus comprised of texts produced in a specialised context of communication, in which the discourse of a community of experts from a field of interest is reflected. The purpose of the creation of this *corpus* is to analyse the discourses of experts in order to extract information that represents the experts' conceptualisations beyond the verbal expression.

The corpus is comprised of 98 texts written in European Portuguese. These 98 texts were produced by experts belonging to different organisations coming from different areas — scientific, industrial, techno-professional, certifying, regulating, and commercial — and are available online. Within this line, we considered specific criteria for the collection of texts to be included in the corpus, focusing on the communication settings of production/reception, where authorship is of utmost importance for the reliability of the information contained in the texts and the intended outcome of the linguistic analysis. The texts were compiled according to the following criteria :

1. texts produced by and for the scientific community of the domain of cork ;
2. texts produced by experts for quasi-experts ;
3. texts produced for non-experts.

The rationale behind the inclusion of the third group in the corpus is the fact that these texts are rich in definitional contexts[6] and/or contexts[7] that describe concepts given the different degrees of knowledge of producers and recipients.

6 By definitional contexts, we mean contexts that are rich in knowledge information permitting the elaboration of definitions.

7 a piece of text that helps to explain the meaning of a linguistic expression.

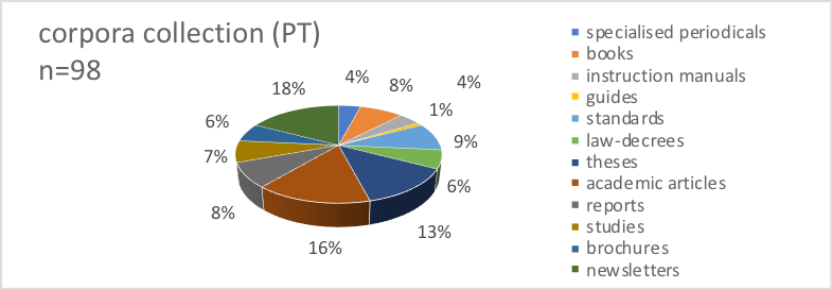


FIG. 3 – *Corpora collection – 98 texts produced following 3 major criteria : expert-expert ; expert-quasi-expert ; expert-non-experts.*

Following the criteria mentioned above, we obtained a balanced corpus that covers the different levels of specialised discourse.

The communicative setting of the production of the texts was the most significant criterion for the compilation of the corpus to support our terminological purposes. An important aspect is that the linguistic analysis was performed on texts produced by experts for semi-experts or quasi-experts that are commonly technical-explanatory, as well as normative texts, and texts produced for the economic and financial areas (the latter were produced by experts of the domain for experts of governmental institutions). The underlying reason for this option is that these texts contain glossaries and definitions produced by experts ; thus, validation of the extracted terms[8] is provided *a priori*. The remaining corpora are used as reference corpora.

4. Collocation

The Anglo-Saxon school (Halliday, 1996 ; Sinclair, 1996 ; Benson, 1988, 1997) and the German school (Hausmann, 1989 ; Heid, 2001) prefer the term collocation to designate sets of units that co-occur in contiguity with a certain frequency in the syntagmatic axis.

In the theoretical framework of Halliday's discourse analysis, the sets of cohesive lexical units deserve attention. For Halliday, cohesion underlies the

8 The analysis of these glossaries and definitions have been at the core of our terminological work since 2013. Thus, a considerable amount of terms and definitions of the domain has already been compiled.

stabilisation of the constituents that co-occur, and a privileged syntactic-semantic relation can be inferred from that co-occurrence:

> Cohesion occurs when the interpretation of some elements in the discourse is dependent on that other. The one presupposes the other, in the sense that it cannot be effectively decoded except by recourse to it (Halliday, 1996, p. 4).

In this theoretical framework, collocations can be analysed from a grammatical or lexical perspective — cohesion, i.e., the essence of collocation as a linguistic entity, is identified by the relationships established between the different lexical units. The lexical units that form a lexical collocation are updated in discourse, because the linguistic system allows for a privileged lexical proximity. This lexical proximity is not always the same:

> There are degrees of proximity in the lexical system, a function of the relative probability with which one word tends to co-occur with another. Secondly, in the text there is relatedness of another kind, relative proximity in the simple sense of the distance separating one item from another, the number of words or clauses or sentences in between (Halliday, 1996, p. 290).

Thus, the density of cohesion is related to how speakers activate the linguistic system and consequently to how they update lexical units, depending on the construction of a discourse or a text.

Sinclair also addresses collocations in their double grammatical and lexical aspects, attributing an essential role to the statistical method for the delimitation of a collocation in order to describe it more adequately:

> Collocation is the occurrence of two or more words within a short space of each other in the text (Sinclair, 1996, p. 170).

Sinclair is particularly concerned with the lexical aspect of collocation, from a lexicographical perspective, focusing on its formal descriptions.

Collocations are thus approached according to lexical and statistical methodologies. Using the concept of distance, understood as the length of the line segment defined by two points, these methodologies allow us to calculate and measure the density between the units that comprise the collocation. The lexical and statistical analyses are, within this scope, complementary tools. The frequency with which each of the units of the collocation is updated in a

given syntagmatic order is an indicator of its identification, description, and classification.

Each of the units that make up the collocation may have a different importance and a different value. The unit under observation consists of a node and a collocate, and any unit may assume the status of node or collocate, depending on the value assigned to each unit.

In 1989, Hausmann defined collocation as :

> […] la combinaison caractéristique de deux mots dans une des structures suivantes : a) substantif + adjectif (épithète); b) substantif + verbe ; c) verbe + substantif ; d) verbe + adverbe ; e) adjectif + adverbe ; f) substantif + (prép.) + substantif (Hausmann, 1989, p. 1010).

Hausmann argues that collocation is defined as opposed to free combination and idiomatic expression by the combinatorial restriction of the units that make up the collocation, and by their transparency and their non-cohesion, being apprehended and used as a unit of language and not as a unit of "parole" in the Saussurean sense. The fact that a collocation can assume the status of lexicographic unit implies going from discourse to language at a given moment.

For Hausmann, a collocation is an oriented combination, i.e., the units that comprise it do not have the same status — one of them is the core that is responsible for the privileged lexical relations that it maintains with its local surroundings. For this reason, it is essential to distinguish the base from the collocate ("Kollokator") — which is equivalent to the notions of node and collocate — since its identification is indispensable for linguistic description and the lexicographic treatment of the collocation, as well as to learn it and subsequently assimilate it. These two elements that constitute the collocation have different partnership statuses, and the treatment to which they are subject, depending on whether the emphasis is placed on the base or the collocate, is also different.

A collocation is thus composed of a base with syntactic and semantic autonomy and a collocate, which adds a characteristic to the base, without modifying its identity.

Heid partially follows Hausmann's reasoning. However, his view on collocation is different because it is terminological. Heid explicitly argues that a collocation may correspond to a term with characteristics and properties that are different from those of terms traditionally identified as compound nouns.

Heid (2001) also refers to the polarity of collocations. A collocation is comprised of two lexemes and potential determinants, quantifiers, and prepositions — one of the lexemes is determined and the other is determinant : these notions correspond to the "node" and "collocate" of Sinclair and the "Basis" and "Kollokator" of Hausmann.

Since his approach to collocation is terminological, one of the lexemes must necessarily be a term, and both lexemes may assume this status. For Heid, from a linguistic point of view, collocations are

> […] a phenomenon of lexical combinatories : they involve the lexical, semantic, and syntactic properties of lexical items and their syntagmatic co-occurrence (Heid, 2001 :788),

which, like linguistic signs, result from a convention.

When he refers to syntactic properties, Heid establishes a relation between collocation and compound word, since he considers that the choice of the components of a compound is determined from the collocational point of view :

> The choice of the components in such noun groups, like the choice of the components of the compounds, is often collocationally determined : there are clear combinatory preferences, often merely conventional, that in many cases go as far as the complete terminological "fixing" of the compounds and noun groups (Heid, 2001, p. 791).

Heid assumes the difficulty of distinguishing collocation from composition based on purely linguistic criteria, knowing that, from the theoretical point of view, it is a very thin line. From a terminological perspective, such a distinction does not prove to be very operational, since in Terminology it is the designation that is at the basis of the identification of the linguistic reality, regardless of the label that is attributed to it :

> From a terminological point of view, we may be more interested in whether the combination of term and collocate can be seen as the denomination of a new concept in its own right (Heid, 2001, p. 791).

With respect to semantic properties, Heid draws on Mel'čuk's theory (1998). Mel'čuk argues that speakers, in the full use of "parole", use collocations to express generic meaning, and accordingly describes collocations starting from the lexical functions that allow him to account for this generic character and which Heid expresses as follows :

> In lexicography, examples of collocations are usually treated in terms of a given collocate with a given base being arbitrary phenomenon that must be memorized (Heid, 2001, p. 793).

That phenomenon also occurs in specialised language, the difference being that the choice of collocate is usually the result of convention and not of free will.

5. Terminological collocation

In terminology, we are interested in introducing a distinction between polylexical terms and terminological collocations. This distinction is fundamental in terminology, because from our point of view these two lexical phenomena should not be confused. In fact, with ISO 1087 : 2019, the term is a *designation* that represents a *general concept* by linguistic means. We would like to stress that from a morphological point of view the term can be monolexical or polylexical.

The terminological collocation is a reality of high relevance in specialised discourse, but unlike terms, it is not defined by conceptual criteria, but by morphological and syntactic criteria, in which a constituent X occurs in a syntagmatic axis in a privileged and frequent way with a constituent Z, and the selection and lexical order is shared by a community of experts. The use of terminological collocations is often evidence of an individual's social and anthropological belonging to a community.

Thus, constituent X corresponds to [V] in our matrix ; constituent Z to [N]. In this paper, [V] corresponds to a common verb that, in the cork processing industry, has acquired a specialised value and governs a noun phrase in which [N] is a specialised term, such as :

1. [[separar]V [as [cortiças]N] SN] VTC
2. [[[escolher]V [as [rolhas de cortiça] N]SN] VTC

The collocation is characterised by being composed of a set of elements, where one of them exerts a morphosyntactic and/or semantic attraction over the other constituents that make up the collocation. In the case of a terminological collocation, one of its constituents is a term that in a syntagmatic context attracts another constituent, which may be terminological or not — the whole of that morphosyntactic construction is a non-term. "On considère un non-terme toute combinatoire lexicale qui, d'un point de vue morphosyntaxique, peut se confondre avec un terme – désignation verbale de concept – mais qui

n'en est pas un, car cette combinatoire lexicale est non désignative." (Costa 2017).

6. Corpus processing : Sketch Engine

In the scope of this research, we have used Sketch Engine to compile, annotate, and query the corpus employing CQL, where RegEx[9] are used. Sketch Engine has an incorporated tagger for Portuguese called FreeLing, which we have used for the queries.

Considering the 98 documents, we have obtained the following quantitative data :

	Frequency
Tokens	1,712,652
Words	1,217,968
Sentences	48,031

Tab. 2 – *Quantitative data regarding the analysed corpus.*

From the observation of the words identified above, we have seen that the most frequent forms that correspond to terms in the domain under analysis are "cortiça" [cork] and "rolha" [stopper].

Forms	**Frequency**	**Percentage per million**
cortiça	16,127	9,416.40
rolha	7,446	4,347.60

Tab. 3 – *Quantitative data regarding "cortiça" and "rolha".*

Considering the frequency of these two terms and consequently the importance they have in the domain under analysis, we shall look at their behaviour in texts.

The FreeLing tagger has certain limitations, as does Sketch Engine itself. FreeLing cannot distinguish between adjectives and past participles, which is, as regards terminological work, highly limiting as observed in Costa (2001): in terms of probability, in Portuguese, the past participle is not usually part of

9 a regular expression is a compact way of describing complex patterns in texts. You can use them to search for patterns and, once found, to modify the patterns in complex ways. They can also be used to launch programmatic actions that depend on pattern. http://gnosis.cx/publish/programming/regular_expressions.html

the morphosyntactic structure of a term, while an adjective can be. This study was performed in the domain of Remote Sensing (Costa, 2001), where that characteristic was observed in the usage of the adjective "colorido" [colourful] as opposed to the usage of the past participle "colorido" [coloured]. This fact causes some noise in the results obtained from CQL queries. On the other hand, Sketch Engine does not allow semantic tagging, which would be a definite plus to retain certain types of forms while rejecting others thus contributing to the reduction of noise in the results obtained.

6.1. Querying the corpus using CQL

We intend to identify verbal terminological collocations (VTC) whose base is the verb "escolher" or "separar" followed by a specialised term used in the domain under analysis, and whose pattern follows a recurring morphosyntactic pattern in Portuguese : [[V] + [N = mono or polylexical term]]$_{VTC.}$ The structure of the term analysed here corresponds to a noun phrase, which, for example, can have the following behaviour :

[N+N] Term
[N+Prep+N] Term
[N+Adj] Term

In the analysed corpus, we have identified structures such as *separar rolhas com defeitos* [V+N= polylexical term] [separating stoppers with defects] or *estabilizar a cortiça cozida* [V+N= polylexical term] [stabilising the boiled cork]. Hence, the first structure shall be segmented as [V +[N+Prep+N] N= Term]$_{VTC}$ and the second one as [V + [N+Adj]N=Term]$_{VTC}$. "Rolhas com defeitos" [stoppers with defects] and "cortiça cozida" [boiled cork] are terms because they designate concepts.

Based on our linguistic knowledge, we employed CQL using regular expressions to identify fundamental lexical-semantic patterns in verbal terminological collocations.

The tags we have used for the CQL constructs are those adopted by Sketch Engine for the Portuguese language: FreeLing part-of-speech tagset[10], a morphological tagger based on EAGLES[11] proposals.

By using Sketch Engine, we are restricted to the part-of-speech tags available of FreeLing in our queries (CQL A to F – see Table 4 below), such as V; VM; VP; D; A and N. Each of these labels have the following value: V=Verb; VM=Main verb; VP=Past Participle; D=Determiner; A=Adjective; and N=Noun. In addition to these, we also used the "character class" [[:punct:]], a RegEx construct, in order to reflect our wish of "no punctuation" in the results (see CQL F, Table 4).

The decision to use the generic tags V=Verb, N=Noun and A=Adjective in the first queries, instead of specific subtypes such as VM=Main Verb or VP=Past Participle, is a consequence of the limitations of FreeLing. We have noticed amongst the results of our CQL and also in the Word Sketch[12] for "rolha" [stopper], that some linguistic forms are either tagged as A or N, such as "técnica" [technical], or tagged as VP instead of A, such as "cobrilhada" [faulty], among others. Therefore, we decided to construe the first CQL in a somewhat non-linguistic sense, but in such a way that one does not mismatch valuable results.

Below we present six possible examples of CQL that correspond to the following combinations:

	Corpus Query Language
CQL A	[tag="V.*"][tag="D.*"]?"cortiça.*|rolha.*" [tag="V.*|N.*"]
CQL B	[tag="V.*"][tag="D.*"]?"cortiça.*|rolha.*"[]{0,2}[tag="V.*|N.*"]
CQL C	[tag="V.*"][tag="D.*"]?"cortiça.*|rolha.*"[]{0,2}[tag="V.*|N.*"]
CQL D	[tag="V.*" & !(tag="V.P.*")] [tag="D.*"]?"cortiça.*|rolha.*" []{0,2} [tag="V.P.*|N.*"]

10 "A tagset is a list of part-of-speech tags (POS tags for short), i.e. labels used to indicate the part of speech and sometimes also other grammatical categories (case, tense etc.) of each token in a text corpus": https://www.sketchengine.eu/portuguese-freeling-part-of-speech-tagset/?highlight=freeling.

11 http://www.ilc.cnr.it/EAGLES/browse.html

12 A summary of a word's behaviour – one of Sketch Engine features.

CQL E	[tag=”VM.*” & !(tag=”V.P.*”) & !(lemma=”ser.*”)] [tag=”D.*”]?”cortiça.*|rolha.*” [tag=”V.P.*|A.*|N.*”]
CQL F	[tag=”VM.*” & !(tag=”V.P.*”) & !(lemma=”ser.*”)] [tag=”D.*”]?”cortiça.*|rolha.*” [word=”.*” & word !=”[[:punct :]]*”]{0,2} [tag=”V.P.*|A.*|N.*”]

Tab. 4 – *CQL queries on the analysed corpus.*

With these six CQL queries we intend to identify the co-texts in which the terms “cortiça” and “rolha” occur, where the obligatory condition is that they are objects of a transitive verb, [V + [N] *polylexical* $_{term}$] VTC, amounting to all the parts of a VTC. Building on the analysis of the results obtained with CQL A, we fine-tuned CQL queries until we obtained CQL F, which allowed us a finer granularity of the results obtained with CQL queries.

6.2. Example : CQL F (cf. Table 4)

CQL F reads as follows, using Python operators incorporated into Sketch Engine :

> Verb (Main transitive verb EXCEPT Past Participle and EXCEPT lemma of the verb to be) + Determinant (or not) + forms started by *cortiça* or *rolha* + [occurrence of zero to 2 forms (any) EXCEPT punctuation] + Verb (any) ONLY Past Participle) or Adjective (any) or Noun (any).

We then applied 4 filters, whose starting point is a transitive main verb followed by a morphosyntactic sequence where the term “cortiça” occurs.

	Corpus Query Language F	**Frequency**
1	tag=”VM.*” & !(tag=”V.P.*”) & !(lemma=”ser.*”)] [tag=”D.*”]”cortiça.*” [word=”.*” & word !=”[[:punct :]]*”]{0,2} [tag=”V.P.|A.*”]	68
2	[tag=”VM.*” & !(tag=”V.P.*”) & !(lemma=”ser.*”)] [tag=”D.*”]”cortiça.*” [word=”.*” & word !=”[[:punct :]]*”]{0,2} [tag=”N.*”]	123

3	[tag="VM.*" & !(tag="V.P.*") & !(lemma="ser.*")] "cortiça.*" [word=".*" & word!="[[:punct:]]*"]{0,2} [tag="V.P.*|A.*"]	58
4	[tag="VM.*" & !(tag="V.P.*") & !(lemma="ser.*")] "cortiça.*" [word=".*" & word!="[[:punct:]]*"]{0,2} [tag="N.*"]	73

Tab. 5 – *Extension of CQL F for the co-text of the term "cortiça".*

After applying all the rules, we obtained the following results:

CQLF	**Good examples**			**Examples on which to decide**		**Bad examples**	
1	separar a cortiça com verde	separando-se a cortiça virgem		golpeia-se a cortiça no sentido vertical	identificar a cortiça com verde	extrair a cortiça com diversos	retirar a cortiça em grandes
2	estabilizar a cortiça após o descortiçamento	Preparar a cortiça para a transformação		identificar a cortiça com verde fresco	retirar a cortiça em grandes pranchas	Estabilizar a cortiça de forma	promover a cortiça e o reforço
3	--	produzir cortiça para utilizações industriais		produzir cortiça de forma sustentável	--	usando cortiça como componentes	
4	extrair cortiça dos ramos	retirar cortiça com maior calibre	tirar cortiça de um sobreiro	--	--	encontrar cortiça com o calibre	tinham cortiça virgem nesse momento

Tab. 6 – *Results obtained applying CQL F and its extensions.*

The decision whether examples are "good examples", "examples on which to decide" or "bad examples" is based on the knowledge that the authors have of the domain and the corpus, as well as on the linguistic knowledge they have regarding word and term formation and the formation of collocations. It should be noted that all the examples shown in Table 6 have not been submitted to expert validation yet. Depending on the expert feedback, data may be reorganised, if necessary.

We repeated the same exercise replacing "cortiça" [cork] with "rolha" [stopper].

Following this CQL, we applied four filters whose starting point is a transitive main verb followed by a morphosyntactic sequence where the term "cortiça" [cork] occurs :

	Corpus Query Language F	Frequency
1	[tag="VM.*" & !(tag="V.P.*") & !(lemma="ser.*")] [tag="D.*"]"rolha.*" [word=".*" & word !="[[:punct :]]*"]{0,2} [tag="V.P.*|A.*"]	142
2	[tag="VM.*" & !(tag="V.P.*") & !(lemma="ser.*")] [tag="D.*"]"rolha.*" [word=".*" & word !="[[:punct :]]*"]{0,2} [tag="N.*"]	286
3	[tag="VM.*" & !(tag="V.P.*") & !(lemma="ser.*")]"rolha.*" [word=".*" & word !="[[:punct :]]*"]{0,2} [tag="V.P.*|A.*"]	61
4	[tag="VM.*" & !(tag="V.P.*") & !(lemma="ser.*")]"rolha.*" [word=".*" & word !="[[:punct :]]*"]{0,2} [tag="N.*"]	71

Tab. 7 – *Extension of CQL F for the co-text of the term "rolha"* [*stopper*].

Applying the rules, we obtained the following results :

CQLF	Good examples			Examples on which to decide		Bad examples	
1	separar as rolhas acabadas	separar as rolhas mal coladas	segregar as rolhas mal colmatadas	--	rastrear as rolhas prontas	classifiquem as rolhas em classes visuais	correspondem a rolhas potenciais foram designadas
2	brocar uma rolha de 24 mm	lavarmos uma rolha de cortiça	--	retirar as rolhas das rabanadas	mergulhar as rolhas com agitação	lavar as rolhas também após colmatagem	transformam as rolhas em bruto

3	--	brocar rolhas naturais	--	produ-zindo rolhas cilíndri-cas	fazer rolhas técnicas	contendo rolhas acabadas prontas	sepa-ram-se as rolhas por classe
4	rectificar rolhas de cortiça natural	brocar rolhas naturais com quali-dade	desin-fectar rolhas de cortiça	--	exporta-vam rol-has dos portos	ver rolhas técnicas	abran-geu rolhas de diversas quali-dades

Tab. 8 – *Results obtained applying CQL F and its extensions.*

According to the examples sampled, we can associate V + term "cortiça" [cork] or "rolha" [stopper] and obtain satisfactory results to identify verbal terminological collocations.

In the following lines, we describe some procedures and operations, which are designated by the terms presented in Fig. 2. The attempt is to demonstrate how these terms occur in specialised texts by means of their morphosyntactic behaviour. In a domain of handicraft and industrial activities, verbs are at the core of the coinage of terms, which accounts for our terminological interest on this subject.

After analysing our corpus, we have identified different verbs to designate the same process. For instance, in the manufacture of the stopper stage, the verbs "escolher" [choose] and "selecionar" [select] are synonyms in the corpus. On the other hand, we also noticed that "escolher" [choose] is replaced by "separar" [separate], both used as synonyms in discourse, although they are only quasi-synonyms since the acts of choosing and separating do not correspond to the same action[13]. We can see that this discursive option introduces ambiguity in the description of the activities of the domain since the verb "separar" [separate] is also used in the debarking stage.

1.CQL: [lemma="separar.*"] []{0,2} "cortiça.*"			
left context	**KWIC**	right context	freq
componente principal	**separa** as cortiças	segundo as	1

13 "separar" [separate]: fazer a disjunção de [divide]; "escolher" [choose]: manifestar preferência por [show preference for] (Dicionário do Houaiss, 2003).

com o intuito de	**separar** porções de cortiça	em grupos,	1
a variável que	**separa** as cortiças	é área mínima	1
ou seja, é	**separada** a cortiça	que não possui	1
as pranchas,	**separar** a cortiça	com verde e	2
,segregadas,	**separadas** da restante cortiça	destinada à	7
poderá ser	**separada** da cortiça	delgada	1
da extracção,	**separando**-se a cortiça	virgem e bocados	1
2.CQL : “cortiça.*” [] {0,2} [lemma=”separar.*”]			
left context	**KWIC**	right context	freq
os grãos de	cortiça, previamente **separados**	em gamas de	1
colhidas amostras de	cortiça e **separadas**	as amostras rolháveis	1
da cozedura, as	cortiças são **separadas**	em fardos de acordo	1
conjunto de pranchas de	cortiça preparada **separadas**	em diferentes classes	3
As pranchas de	cortiça devem ser **separadas**	do solo por	3
impacto dos martelos na	cortiça (**separam**	a lenha da cortiça)	1
3.CQL : [lemma=”seleccionar.*”] []{0,2} “cortiça.*”			
left context	**KWIC**	right context	freq
cortiça cuidadosamente	**selecionados** aglutinados com cortiça	. São diversas as	1
4.CQL : “cortiça.*” [] {0,2} [lemma=”selecc?ionar.*”]			
left context	**KWIC**	right context	freq
o crescimento da	cortiça **seleccionam-se**	da bibliografia	1
para o crescimento	cortiça foi **seleccionado**	tendo em conta	1
conjunto de dez	cortiças **seleccionadas**	na oficina do	1
as pranchas de	cortiças são **seleccionadas**	de acordo com a	1
com granulados de	cortiça cuidadosamente **seleccionados**	e aglutinados com borracha	1
O pavimento em	cortiça foi **seleccionado**	para responder às	1

com granulados de	cortiça cuidadosamente **seleccionados**	em que o látex	1
com granulados de	cortiça cuidadosamente **seleccionados**	aglutinados com cortiça	1
dois ou três discos de	cortiça natural **seleccionada**	As rolhas aglomeradas	2
Passado este período ,	cortiça é então **seleccionada**	nomeadamente no que	2
5.CQL : [lemma=”escolher.*”][]{0,2}”cortiça.*”			
left context	**KWIC**	right context	freq
preparadora. Esta indústria	**escolhe** as cortiças	empilhadas de acordo	1
Herdade de Espirra foram	**escolhidas** as cortiças	em função do calibre	1
monge beneditino,	**escolheu** as rolhas de cortiça	para vedar o seu famoso	1
6.CQL : “cortiça.*”[]{0,2}[lemma=”escolher.*”]			
left context	**KWIC**	right context	freq
as pranchas de	cortiça amadia, **escolhe**	novamente por qualidades	1
ou seja, das	cortiças **escolhidas**	e classificadas “(	1
7.CQL : [lemma= “separar.* “] []{0,2} “rolha.*”			
left context	**KWIC**	right context	freq
operação destinada a	**separar** as rolhas	acabadas em classes	3
componente principal	**separa** as rolhas	que apresentam valores	1
, ou seja,	**separam**-se as rolhas	por classe e com defeito	1
Operação destinada a	**separar** as rolhas	em determinado número	20
encontrar-se fisicamente	**separadas** das rolhas	e dos discos,	1
obrigatórias : 5.3.1	**Separar** as rolhas	em função das referências	1
devem estar fisicamente	**separadas** das rolhas	não lavadas, quer	1
que se destina a	**separar** as rolhas	com defeitos de colagem	3
4.2 Objectivo :	**Separar** as rolhas	mal coladas	2
que se destina a	**separar** as rolhas	com defeitos 3.2	4
que consiste em	**separar** as rolhas	ou discos em várias categorias	1

eu aspecto visual e /ou	**separar** as rolhas	com defeitos 2.3	4
5.2 Objectivo :	**Separar** as rolhas	mal coladas.	1
8.CQL : “rolha.*” [] {0,2} [lemma=”separar.*”]			
left context	**KWIC**	right context	freq
programadas e as	rolhas são **separadas**	, com um mecanismo	1
espumantes . 922 . Estas	rolhas estão geralmente **separadas**	em classes “Extra”,	1
imperfeições que as	rolhas apresentem, **separando**-as	concomitantemente, em classes	1
9.CQL : [lemma=”seleccionar.*”] []{0,2} “rolha.*”			
left context	**KWIC**	right context	freq
de cortiça natural	**seleccionada** . As rolhas	aglomeradas são inteiramente	2
10.CQL : “rolha.*” [] {0,2} [lemma=”selecc ?ionar.*”]			
left context	**KWIC**	right context	freq
Selecção : processo no qual as	rolhas são **seleccionadas**	de acordo com a sua qualidade	1
O comprimento da	rolha **seleccionada**	deve estar de acordo	1
11.CQL : [lemma=”escolher.*”][]{0,2}”rolha.*”			
left context	**KWIC**	right context	freq
monge beneditino,	**escolheu** rolhas	de cortiça para vedar	1
normal) , devem-se	**escolher** rolhas	com um diâmetro superior	2
12.CQL : “rolha.*”[]{0,2}[lemma=”escolher.*”]			
left context	KWIC	right context	freq
qualidade associado. As	rolhas depois de **escolhidas**	separadas seguem para	1
- Escolha visual As	rolhas são **escolhidas**	em máquinas electrónicas	1

Tab. *9 – Verbs “separar”[separate]; “seleccionar” [select]; and “escolher” [choose] in co-text with “cortiça” [cork] or “rolha” [stopper]. KWIC were drawn from the cork corpus with CQL queries.*

Table 9 contains the verbs “separar” [separate] and “escolher” [choose] in co-occurrence with “cortiça” [cork], obtained from the cork corpus using CQL interrogations. The verbs that are the starting point in the CQL are highlighted, either on the left or right-side of the key word in context (KWIC).

We can observe that both "cortiça" [cork] and "rolha" [stopper] widely co-occur with several inflexions of the verb "separar" [separate] (e.g., *separa as cortiças; separada a cortiça; separar a cortiça / separam-se as rolhas; separar as rolhas; separadas das rolhas*). However, while "cortiça" [cork] has a high co-occurrence with "seleccionar" [select], "rolha" [stopper] has a very low co-occurrence with this verb, as seen on CQL 3. and 4. Vs. CQL 9. and 10. Finally, the verb "escolher" [choose] has a shred of minor evidence for both "cortiça" [cork] and "rolha" [stopper], as shown on CQL 5.; 6.; 11; and 12.

7. Conclusions

The purpose of this research was to prove that verbal terminological collocations are linguistic structures that, together with polylexical terms, play a fundamental role in expert discourse. However, they perform different functions: although they may have morphosyntactic and lexical structures that are actually the same or similar, polylexical terms and terminological collocations are distinguished by the criteria underlying the analysis: terms are governed primarily by conceptual criteria and collocations by morphosyntactic criteria.

The analysis we have carried out in this paper aims to demonstrate how morphosyntactic analysis is complementary to a more concept-focused analysis, allowing us to obtain information that can feed different terminological resources (dictionaries, ontologies, …).

In the domain of the cork industry, common verbs in Portuguese acquire specific meaning when occurring in co-text with terms; an evidence observed through the analysis of the recursive morphosyntactic constructions found in the corpus. These structures underpin our distinction of a polylexical term from a verbal terminological collocation.

This paper had three purposes that the authors believe to be fundamental for the terminological work:

1. Associating domain knowledge and the linguistic analysis of how texts work;

2. Based on that knowledge, creating local grammars from the analysis of co-texts, in this case, for transitive verbs;

3. Using text mining tools to increase knowledge on the behaviour of the combinations under analysis;

4. Including data validation criteria.

Automatic language processing tools have their limitations. Sketch Engine is no different. Some of the bad results obtained are originated by FreeLing limitations, which forces the user to be somewhat creative in order to capture any meaningful silence and/or eliminate noise. Using this methodology, text processing is still an overly labour-intensive and time-consuming task.

The work that has been carried out at NOVA CLUNL since 2001 is now being updated so it contains semantic information that will increase the quality of the data obtained.

Acknowledgements : Research financed by Portuguese National Funding through the FCT – Fundação para a Ciência e Tecnologia as part of the project Centro de Linguística da Universidade NOVA de Lisboa – UID/LIN/03213/2020 and the PhD program in Linguistic KRUse - Knowledge, Representation & Use, CLUNL - Faculty of Social Sciences and Humanities, Universidade NOVA de Lisboa - PB/BD/113972/2015.

References

Benson, Morton, Benson, Evelyn and Ilson, Robert. 1988. "The BBI Combinatory Dictionary of English. A Guide to Word Combinations". In *Revue belge de philologie et d'histoire,* Vol. 66 No. 3, 709-710. Langues et littératures modernes - Moderne taal- en letterkunde.

Benson, Morton, Benson, Evelyn and Ilson, Robert. 1997. "The BBI Dictionary of English Word Combinations.", VII - XXXIX. Amsterdam, Philadelphia : John Benjamins

Costa, Rute. 2017. "Les collocations terminologiques." Provas de agregação, Lexicologia, Lexicografia, Terminologia. Lisbon : FCSH UNL.

Costa Rute. 2013. "Terminology and Specialised Lexicography : two complementary domains". In *Lexicographica. International Annual of Lexicography,* Vol. 29 No. 1, edited by Gouws, Rufus Hjalmar / Heid, Ulrich / Schierholz, Stefan J. / Schweickard, Wolfgand / Wiegand, Heribert Ernst. Berlin, New York : De Gruyter.

Costa, Rute. 2001. "Pressupostos teóricos e metodológicos para a extracção automática de unidades terminológicas multilexémicas". PhD dissert., Lisbon : FCSH UNL.

Halliday, M. A. K..1991. "Corpus Studies and Probabilistic Grammar". In *English Corpus Linguistic, Studies in Honour of Jan Svartvik,* edited by Karin Aijmer & Bengt Altenberg, 30 - 43. London, New York : Longman.

Hausmann, Franz Josef. 1989. “Le dictionnaire des collocations”. In *Wörterbucher, Ein internationales Handbuch fur Lexicographie,* edited by Franz Josef Hausmann, Oskar Reichmann, Herbert Ernst Wiegand, Ladislav Zgusta, 1010 - 1019. Berlin, New York : Walter de Gruyter.

Heid, Ulrich. 2001. “Collocations in Sublanguage Texts : Extraction from Corpora.” In *Handbook of Terminology Management. Application-Oriented terminology Management,* Vol. 2, compiled by Sue Ellen Wright and Gerhard Budin, 788 - 808. Amsterdam, Philadelphia : John Benjamins Publishing Company.

ISO 1087-1. 2019. “Terminology work and terminology science — Vocabulary”. Genève : Organisation Internationale de Normalisation.

Mel’cuk UK, Igor A..1998. “Collocations and Lexical Functions, Phraseology, Theory, Analysis, and Applications”, edited by A. P. Cowie, 23 – 54. Oxford : Oxford University Press.

Nunes, Paulo. 2013. “Análise do fluxo de processo industrial e do respetivo plano de inspeção e ensaios.” Ma dissert., Porto : FEUP Universidade do Porto.

Ramos, Margarida and Costa, Rute. 2018. “Semantic Analyses of Texts for Eliciting and Representing Concepts : the TermCork Project.” In *Actes de la dixième conférence TOTh 2016, 9-10 June.* Chambéry : Institut Porphyre, Savoie et Connaissance.

Ramzan, Talib ; Muhammad, K. Hanify ; Shaeela, Ayeshaz & Fakeeha, Fatima. 2016. “Text Mining : Techniques, Applications and Issues”. In *(IJACSA) International Journal of Advanced Computer Science and Applications,* Vol. 7 No. 11, 414 – 418. Available at https ://thesai.org/Publications/IJACSA.

Silva, Raquel & Costa, Rute. 2019. “Accéder aux connaissances des experts par l’entremise de la médiation en terminologie”. In *L’essentiel de la médiation. Le regard des sciences humaines et sociales,* edited by Michele De Gioia and Mario Marcon, 105 – 121. Bruxelles, Bern, Berlin, New York, Oxford, Wien : P.I.E. Peter Lang

Sinclair, John : Ball, J. 1996. “EAGLES : Preliminary Recommendations on Text Typology (EAG - TCWG - TTYP/P).”, Version of June, pp. 71. Available at http://www.ilc.pi.cnr.it.

Résumé

Le but de cet article est d’organiser les connaissances lexicales et conceptuelles en analysant un corpus spécifique à un domaine. Le domaine sur lequel

nous nous concentrons est l'industrie du liège. Grâce à l'analyse du corpus, nous avons constaté que certains verbes communs en portugais, tels que « choisir » et « séparer » acquièrent une valeur spécialisée dans le domaine à l'étude. Ce fut le point de départ de l'analyse des collocations terminologiques où les verbes sont les constituants centraux, dans la perspective de la double dimension de la terminologie. Pour l'analyse de ces collocations terminologiques verbales, nous avons utilisé des techniques de traitement du langage naturel dans lesquelles des structures CQL simples à plus complexes ont été construites avec REGEX. Le résultat de cette analyse nous permet d'introduire une distinction entre termes polylexicaux et collocations terminologiques. La collocation terminologique est une réalité d'une grande pertinence dans le discours spécialisé, mais contrairement aux termes, elle n'est pas définie par des critères conceptuels, mais par des critères morphologiques et syntaxiques.

Articles session « ELEXIS »

Using an Infrastructure for Lexicography in the Field of Terminology

Tanja Wissik*, Thierry Declerck*/**

*Austrian Academy of Sciences
Austrian Centre for Digital Humanities and Cultural Heritage
tanja.wissik@oeaw.ac.at
https://www.oeaw.ac.at/acdh/team/current-team/tanja-wissik
**DFKI GmbH
Multilinguality and Language Technology Lab
Stuhlsatzenhausweg, 3
D-66123 Saarbrücken
Germany
declerck@dfki.de
https://www.dfki.de/~declerck/

Abstract. In this contribution, we discuss the (re-)use of the ELEXIS research infrastructure for lexicography in order to deal with terminological data. We present central aspects of the ELEXIS infrastructure and the standards it both applies and further develops. We also present TBX, which is the main standard used for representing terminological data. We describe in some detail the OntoLex-Lemon specifications, which which result from a W3C Community Group and which play a central role in our work consisting in describing terminological data within an infrastructure for lexicography, as it supports linking knowledge organisation systems to a full lexical description. To exemplify this capability, we use multilingual terminology data, originally encoded in TBX, from the field of risk management.

1. Introduction

In many disciplines e-Research and the use of digital methods has become an omni-present research practice (Lusicky and Wissik (2017)). In general, research infrastructures enable e-Research by providing facilities, resources or services of a unique nature, to conduct and to support top-level research

activities in different domains. They include for example major scientific equipment or sets of instruments, knowledge-based resources like collections, archives and scientific data as well as e-Infrastructures, such as data and computing systems and communication networks, and any other tools that are essential to achieve excellence in research and innovation.[1]

2. Research Infrastructures

Research infrastructures (RIs) offer technical and social infrastructures in a more stable and sustainable way than research projects that run only for two to four years. As technical infrastructure they provide resources, tools and services to the scientific community in order to support top-level research activities. As social infrastructure, RIs provide platforms for collaborative research and knowledge transfer and promote the use of common methods and standards. They also play an important role in educating new generations of researchers.

The European Strategy Forum on Research Infrastructures (ESFRI) recognises over fifty different research infrastructures (ESFRI (2018)). While some infastructures are domain specific and others generic, ideally these domain specific and generic research infrastructures complement each other (Illmayer (2017)).

2.1. Generic Research Infrastructures

As Generic Research Infrastructures, for this paper, we understand Research Infrastructures that can be used by researchers from a variety of research fields. In the Humanities, e.g. CLARIN (Common Language Resources and Technology Infrastructure) and DARIAH (Digital Research Infrastructure for the Arts and Humanities) count as generic research infrastructures (Doel and Maes (2012)).

2.2. Domain Specific Research Infrastructures

Besides generic research infrastructures, there are domain specific research infrastructures. In the Humanities, e.g. EHRI (European Holocaust Research Infrastructure) or ARIADNE (Advanced Research Infrastructure

1 ESFRI Roadmap 2018, Part 1, p 11

for Archaeological Dataset Networking in Europe) can be seen as domain specific research infrastructure.

3. European Lexicographic Infrastructure (ELEXIS)

As an example of a domain specific research infrastructure we introduce the European Lexicographic Infrastructure (ELEXIS). ELEXIS is a research infrastructure project under H2020 aiming to develop a Research Infrastructure for Lexicography that provides online access to data, tools and services for lexicography research (cf. Declerck *et al*. 2018). In the following we describe the whole infrastructure, also the parts that are still in development. It consists of three sub-infrastructures LEX1, LEX2, LEX3 (Krek *et al*. (2018)) that are explained below in more detail.

The first part of the infrastructure (LEX1) includes conversion and alignment tools in order to harmonise and convert lexicographic resources into a uniform data format that allows their integration in the Linked Open Data (LOD) cloud[2], and more specifically in the Linguistic Linked Open Data (LLOD) cloud[3]. The LLOD was originally an initiative by members of the Open Knowledge Foundation that has gained a lot of attraction, and which was further developed in various projects, as described in (McCrae *et al*. (2016)). The second part of the infrastructure (LEX2) includes word sense disambiguation and entity linking tools dedicated to semantic processing of corpus data. These tools facilitate disambiguation and corpus analysis and open up the possibility to create lexicographic resources automatically from corpora. The third part of the infrastructure (LEX3) includes tools to support the retro-digitising process of dictionaries (Krek *et al*. (2018)).

As common data formats, the ELEXIS infrastructure makes use of the OntoLex-Lemon (McCrae *et al*. (2017)) and TEI Lex-0 models (Banski *et al*. (2017)) as outlined in (Ahmadi *et al*. (2019)) or in (McCrae *et al*. (2019)).

There are different ways, how the ELEXIS Infrastructure can be used for already existing dictionary data. For dictionaries that are not already in a digital format the retro-digitization tool in LEX3 is used. This applies OCR to the text and then processes it by adding XML markup in the format of TEI Lex-0. For dictionaries that are already available in a digital form, but not one that is supported directly by the project, the conversion tool of LEX1 is

2 See https://www.lod-cloud.net/ for more details.
3 See https://linguistic-lod.org for more details.

used to convert these resources to TEI Lex-0. If the dictionaries are already in TEI Lex-0 or have been converted to TEI Lex-0 by one of the two methods described above, they can be consumed directly by the interoperable REST interface. Furthermore, if the dictionaries are already available in Ontolex-Lemon, they can also be consumed directly be the REST interface. However, there are no conversion tools foreseen in the infrastructure to convert other formats to Ontolex-Lemon, for the time being it is handled outside of the infrastructure. Most of this mapping work onto OntoLex-Lemon is realised by a partner Research and Innovation project, Prêt-à-LLOD[4]. One of its tasks is to transform various types of linguistic data into an LLOD compliant format.

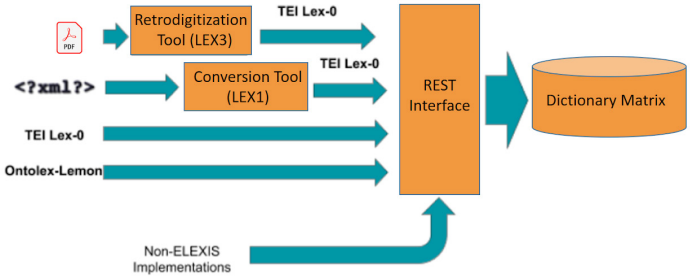


Fig. 1 – *Used formats and standards in order to access the ELEXIS Infrastructure via the REST Interface (Ahmadi et al., 2019, modified).*

All data, copyrights permitting, will be made available as part of the Linguistic Linked Open Data cloud.

In the next section we discuss in detail the different standards used for terminological and lexicographical resources, especially those relevant for the ELEXIS infrastructure and our use case.

4. Standards

One aspect of a sustainable infrastructure and the (re)-usability of resources, services and tools are common standards and data formats. In the

4 See https://www.pret-a-llod.eu/ for more details.

following section we describe standards that are relevant in the fields of lexicography and terminology.

4.1. Standards for Lexicography

In terms of used standards and formats, the field of lexicography is very heterogeneous. In the ELEXIS project, a survey of user needs was carried out (Kallas *et al*. (2019)). This survey had two parts, one for lexicographers and one for institutions. In the survey, there were also questions regarding used standards and formats. The survey has shown that many lexicographic projects use XML or databases and some RDF based formats, but there are still projects working with unstructured data and text format. Among those using XML, custom XML, TEI (P2 or P5) and TEI Lex-0 were mentioned (Kallas *et al*. (2019)). According to the survey data, two tendencies were observed: "a) a transition from non-structured data or text format to structured data format; b) still insufficient use of (standardised) structured formats enabling reliable re-use and linking of dictionary data" (Kallas (2019, 55)). In this respect, ELEXIS, as a promoter of common standards, plays a crucial role. In the following we describe some of the standards mentioned in the ELEXIS survey of user needs.

4.1.1. TEI Dictionary Chapter

In the TEI P5 Guidelines, the Dictionary Chapter defines how to encode lexical resources of all kinds, in particular human-oriented monolingual and multilingual dictionaries, glossaries, and similar documents (TEI P5 2019).

4.1.2. TEI Lex-0

The TEI Dictionary Chapter is very complex and allows many ways to encode a dictionary entry. TEI Lex-0 is simpler and aims at establishing a baseline encoding and a target format to facilitate the interoperability of heterogeneously encoded lexical resources (DARIAH WG Lexical Resources (2019)). TEI Lex-0 should not be thought of as a replacement of the Dictionary Chapter in the TEI Guidelines or as the format that must be used for editing or managing individual resources, especially in those projects and/or institutions that already have established workflows based on their own flavors of TEI. TEI Lex-0 should be primarily seen as a format that existing TEI dictionaries can be univocally transformed to in order to be queried, visualised, or mined in a uniform way (Romary (2015), DARIAH WG Lexical Resources (2019)).

4.1.3. OntoLex-Lemon

The OntoLex-Lemon model was originally developed with the aim to provide a rich linguistic grounding for ontologies, meaning that the natural language expressions used in the labels, definitions or comments of ontology elements are equipped with an extensive linguistic description.[5] This rich linguistic grounding includes the representation of morphological and syntactic properties of lexical entries as well as the syntax-semantics interface, i.e. the meaning of these lexical entries with respect to an ontology or to specialised vocabularies.

The main organising unit for those linguistic descriptions is the lexical entry, which enables the representation of morphological patterns for each entry (a multi word expression, a word or an affix). The connection of a lexical entry to an ontological entity is marked mainly by the *ontolex:denotes* property or is mediated by the *LexicalSense* or the *LexicalConcept* classes, as this is represented in Figure 2, which displays the core module of the model. OntoLex-Lemon builds on and extends the lemon model (Cimiano *et al*. (2016)). A major difference is that OntoLex-Lemon includes an explicit way to encode conceptual hierarchies, using the SKOS[6] standard. As can be seen in Figure 2, lexical entries can be linked, via the *ontolex:evokes* property, to such SKOS concepts, which can represent WordNet synsets. This structure is paralleling the relation between lexical entries and ontological resources, which is implemented either directly by the *ontolex:reference* property or mediated by the instances of the *ontolex:LexicalSense* class.

5 See (McCrae *et al*., 2012), (Cimiano *et al*., 2016) and also https://www.w3.org/2016/05/ontolex/

6 SKOS stands for "Simple Knowledge Organization System". SKOS provides "a model for expressing the basic structure and content of concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies, folksonomies, and other similar types of controlled vocabulary" (https://www.w3.org/TR/skos-primer/).

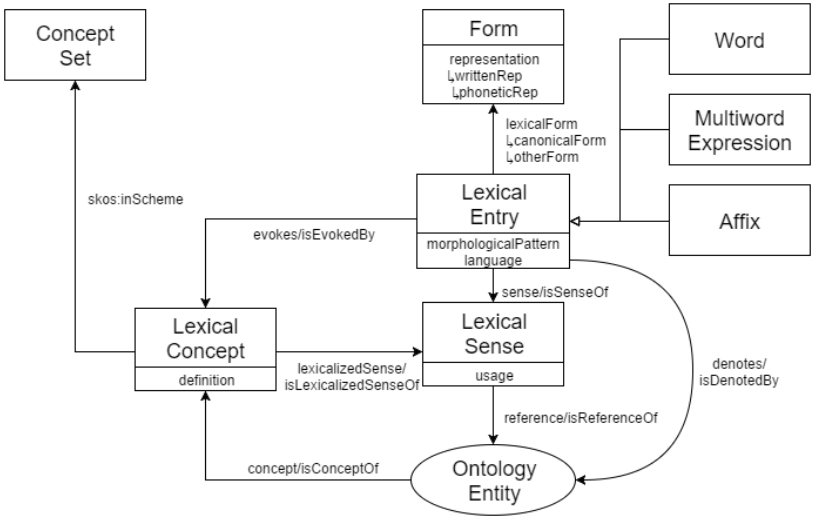


Fig. 2 – *The core module of OntoLex-Lemon : Ontology Lexicon Interface. Graphic taken from https ://www.w3.org/2016/05/ontolex/.*

More recent developments of the model have been described in (McCrae *et al*. (2017)). Currently three extension modules are being discussed : a lexicographic, a morphology and a corpus frequency module[7].

4.2. Standards for Terminology

Most standards in the field of terminology are language technology related standards. One of these standards is TBX, or TermBase eXchange, the other is TBX basic, a reduced version of TBX.

4.2.1. TBX

TBX is the international standard for representing and exchanging information on terminological data. It defines a family of formats that share a common structure and a limited range of information types. Each member of the family is called a dialect of TBX. The main purpose of TBX is to ensure "the independence of valuable terminological data from any particular software

7 See respectively https ://www.w3.org/2019/09/lexicog/, https ://www.w3.org/community/ontolex/wiki/Morphology and https ://github.com/acoli-repo/ontolex-frac

application used to access, display, update or otherwise process it" (Melby 2015, 393).

TBX is modular in order to support the varying types of terminological data, or data-categories, that are included in different terminological databases (termbases). TBX includes two modules: a core structure, and a formalism for identifying a set of data-categories and their constraints, both expressed in XML. The term TBX, when used alone, refers to the framework consisting of these two interacting modules.

The TBX entry model is organised in strict compliance to ISO 16642 (TMF) in that each of three elements: <termEntry>, <langSet> and <tig> in TBX respectively correspond to the three levels in TMF (entry level, language level and term level) (cf. Romary 2014). The data categories associated to these three levels are made of a) specific elements such as <term>, <note>, <ref> and <xref> and b) so called meta data-elements that may express a wide range of possible data categories, namely <admin>, <descrip> and <termNote>. For instance, <descrip type= "definition"> is how a definition is represented in TBX. (Romary (2014)).

4.2.2. TBX basic

TBX basic, as already mentioned, is a simpler and reduced version of TBX. It is also XML based as TBX and adheres to the same entry structure as described above (entry level, language level and term level) but only allows a limited set of data categories. Its purpose is to formalise the markup that is used in relatively simple terminology resources, in order to ensure interoperability (Terminorgs (2014)).

5. Use Case

5.1. Resource Description

The data we use for this use case is a data export of a multilingual terminology database on risk management provided by the University of Vienna available in the ELRC-SHARE repository[8]. The resource is in TBX-Basic format containing 1024 terms in the field of risk management in 5 languages: French, English, German, Romanian and Spanish. The termbase was created

8 https://www.elrc-share.eu/repository

in order to improve domain communication and to facilitate mutual understanding across linguistic boundaries. The intended target users of this terminology resource were risk managers, civil engineers as well as teachers, students and translators (cf. Budin (2011, 23)).

The term entries contain terms, definitions (sometimes even more than one definition is provided) and context information, but for example, no grammatical information like part of speech is available as a separate data category. The resource contains single word expressions such as "risk" or "riesgo" as well as multiword expressions such as "risk reduction" or "reducción del riesgo" or "disaster risk reduction" or "reducción del riesgo de desastre".

5.2. TBX to Ontolex Lemon Transformation

So far there have been no mapping efforts to map common formats in terminology such as TBX to TEI Lex-0. However, there have been discussions on how to provide a representation of onomasiological data, such as terminological data in TEI in addition to the already existing "dictionaries" chapter (Romary (2014)). Furthermore, there have been initiatives to describe terminological data in RDF based representations (Cimiano *et al*. (2015), Rodriguez-Doncel *et al*. (2018)). From there it is possible to convert already existing terminological data to OntoLex-Lemon, which is the core of the LEX1 component of ELEXIS. As mentioned earlier, OntoLex-Lemon provides for a declarative interface between knowledge systems represented in SKOS and lexical data represented in OntoLex. In this, one can easily combine terminologies and lexicographical data and make them interoperable.

For exemplifying our approach, we display below some code from the OntoLex-Lemon encoding, in a simplified form, of one term taken from the Risk Management Terminology. The term is in the original TBX "risk". While the original terminology is repeating the term by each covered language ("risk", "risque", "Risiko", "Risc", "riesgo", for EN, FR, DE, RO and ES), in our conversion we have only one SKOS concept. We also note that in the original terminology the terms in the various languages are considering some typographical rules (for example capital letter for German nouns). We advocate for a language independent and neutral encoding of the terms of a termbase. This way, we implement a modular approach in which the terms are organised independently of the language data they use, but to which they are inter-linked by declarative property relations.

For reason of space and of simplicity we give just an example of the term "risk" below, but the OntoLex-Lemon model can also deal with terms involving multy-word expressions.

```
risk :ConceptSet_1
a ontolex :ConceptSet ;
rdfs :label "Risk Management Terminology from the University of
Vienna"@en.
```

The first code example displays the introduction of an instance of a skos :ConceptScheme (see Fig. 2 as orientation for all displayed code examples)

```
risk :LexicalConcept_1
a ontolex :LexicalConcept ;
rdfs :label „risk"@en ;
rdfs :label „riesgo"@es ;
rdfs :label „risque"@fr ;
rdfs :label „Risl"@de ;
skos :definition "Probabilidad de que un evento ocurra. Cálculo
matemático de pérdidas (de vidas, personas heridas, propiedad dañada y
actividad económica detenida) durante un periodo de referencia en una
región dada para un peligro en particular. Riesgo es el producto de la
amenaza y la vulnerabilidad."@es ;
skos :topConceptOf risk :ConceptSet_1 ;
ontolex :isConceptOf <https ://www.wikidata.org/wiki/Q104493> ;
ontolex :isEvokedBy risk :Word_1 .
```

Now we introduce the term that will cover all the language variations associated with it (so that we do not duplicate anymore the number of terms for one concept). In the code example before, we added just one definition, the Spanish one. The instance of *LexicalConcept* is related to a *LexicalEntry* by the property *isEvokedBy* and to an ontology entry in Wkidata by the property *isConceptOf*. Lexical data is thus no longer encoded with the conceptual space but linked to it.

The following code example is displaying the basic lexical information we are associated with the term "LexicalConcept_1" (to which it relates with the property "evokes". We specify there not only the language of the entry but also its part-of-speech and its gender, information which is not included in the original terminology.

```
risk :Word_1
a ontolex :Word ;
```

```
dc :language “http://id.loc.gov/vocabulary/iso639-2/spa” ;
lexinfo :gender lexinfo :masculine ;
lexinfo :partOfSpeech lexinfo :noun ;
rdfs :label “riesgo”@es ;
ontolex :canonicalForm risk :Form_1 ;
ontolex :evokes risk :LexicalConcept_1 .
```

The next two code examples are encoding the morphological variants of the lexical entry. This is an important information as some terms are restricting the usage of some words to be either singular or plural. The above mentioned OntoLex-Lemon module “lexicog” is specifying the way such usage restrictions can be encoded.

```
risk :Form_1
a ontolex :Form ;
lexinfo :number lexinfo :singular ;
ontolex :writtenRep “riesgo[@es}” .
risk :Form_2
a ontolex :Form ;
ontolex :representation “riesgos”@es .
```

These code examples show only a simple term, consisting of only a unique word. But the OntoLex-Lemon supports also the linking of terms consisting of multi word expressions to a full description of the lexical units of the term.[9] We note that the mapping from TBX to OntoLex-Lemon is not just realising a format conversion, but it is also leading to a simplification of the the original terminological data, as it does not need to include lexical information any more but can link to a specialised lexical data set. Another important aspect is the fact that we are now able to express lexical restriction on the words used in the terms.

6. Conclusion

In this paper, we outlined the possibility to use a domain specific infrastructure for another domain, for which it was not initially created, and we identified the key points for the usage. As use case we have chosen a research infrastructure for lexicography, showing how it can be used in the field of terminology. In the described use case, the key points to be able to use the

9 (Tiberius and Declerck (2017)) describe how Dutch compounds can be represented in their constituent parts. See also for the relevant sections of https ://www.w3.org/2016/05/ontolex/.

infrastructure are common formats and standards or the possibility of mapping and converting them, and supporting interoperability between different types of language data : terminological and lexical data.

Acknowledgements. Contributions by the Austrian Centre for Digital Humanities and Cultural Heritage at the Austrian Academy of Sciences were supported in part by the H2020 project "ELEXIS" with Grant Agreement number 731015. Contributions by the German Research Center for Artificial Intelligence (DFKI GmbH) were supported in part by the H2020 project Prêt-à-LLOD with Grant Agreement number 825182.

References

Ahmadi, Sina, Arcan, Mihael, Declerck, Thierry, Kernerman, Ilan, Khan, Fahad, Krek, Simon, McCrae, John, Měchura, Michal, Monachini, Monica, Roche, Christophe, Tiberius, Carole, Troelsgård, Thomas, Zaytseva, Ksenia. 2019. D2.1.Interface for Interoperable Lexical Resources. Accessed September 19. https ://elex.is/wp-content/uploads/2019/02/ELEXIS_D2_1_Interface_for_Interoperable_Resources.pdf

Bański, Piotr, Bowers, Jack, Erjavec,Tomaz. TEI-Lex0 guidelines for the encoding of dictionary information on written and spoken forms. Electronic Lexicography in the 21st Century : Proceedings of eLex 2017 Conference, Sep 2017, Leiden, Netherlands. <hal-01757108>

Budin, Gerhard. 2011. Designing and Implementing Strategies of Global, Multilingual "Disaster Communication". In Alekseeva, Larissa (ed). The 18th European Symposium on Language for Special Purposes : Proceedings. Perm State National Research University, Perm, Russia. 11-26.

Cimiano, Philipp, McCrae, John and Paul Buitelaar. 2016. Lexicon Model for Ontologies : W3C Community Report.

Cimiano, Philipp, McCrae, John, Rodriguez-Doncel, Victor, Gornostaya, Tatiana, Gomez-Perez, Asuncion, Siemoneit, Benjamin, Lagzdins, Andis. 2015. Linked Terminology : Apply-ing Linked Data Principles to Terminological Resources. Proceedings of eLex 2015. 504-517.

DARIAH WG Lexical Resources. 2019. TEI Lex-0 — A baseline encoding for lexicographic data. Accessed September 19 https ://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html#index.xml-body.1_div.1_div.1

Declerck, Thierry, McCrae, John, Navigli, Roberto, Zaytseva, Ksenia, Wissik, Tanja . 2018. ELEXIS - European Lexicographic Infrastructure :

Contributions to and from the Linguistic Linked Open Data : Proceedings of the 2nd GLOBALEX Workshop, Japan.

Doel, Wim van den and Katrien Maes. 2012. Social Sciences and Humanities : Essential Fields for European Research and in Horizon 2020. League of European Research. Accessed November 14. https ://www.leru.org/files/Social-Sciences-and-Humanities-Essential-Fields-for-European-Research-and-Horizon-2020-Full-paper.pdf

ESFRI. 2018. "ESFRI roadmap 2018". Accessed June 19. http://roadmap2018.esfri.eu/media/1066/esfri-roadmap-2018.pdf

Illmayer, Klaus. 2017. Aufbau einer digitalen Infrastruktur für Theaterwissenschaft. Skizze einer digitalen Forschungsplattform. Presentation at DHA 2017, Innsbruck Austria. https ://zenodo.org/record/1123312#.XF1Q35rA9PY

Kallas, Jelena, Koeva, Svetla, Kosem,Iztok,Langemets, Margit, Tiberius, Carole. 2019. D1.1. Lexicographic Practices in Europe : A Survey of User Needs. Accessed September 19. https ://elex.is/wpcontent/uploads/2019/02/ELEXIS_D1_1_Lexicographic_Practices_in_Europe_A_Survey_of_User_Needs.pdf

Krek, Simon, Iztok Kosem, John P. McCrae, Roberto Navigli, Bolette S. Pedersen, Carole Tiberius, Tanja Wissik. 2018. European Lexicographic Infrastructure (ELEXIS). In Proceedings of the XVIII EURALEX International Congress Lexicography in Global Contexts. 881-891.

Lušicky, Vesna &Wissik. Tanja. 2017. Discovering Resources in the VLO : A Pilot Study with Students of Translation Studies. In Selected papers from the CLARIN Annual Conference 2016, Aix-en-Provence, 26-28 October 2016, CLARIN Common Language Resources and Technology Infrastructure, ed. L. Borin, 63-75. Linköping University Electronic Press, Linköpings universitet.

McCrae, John, Aguado-de Cea, Guadalupe, Buitelaar, Paul, Cimiano, Philipp, Declerck, Thierry, Gomez-Perez, Asuncion, Garcia, Jorge, Hollink, Laura, Montiel-Ponsoda, Elena, Spohr, Dennis and Wunner, Tobias. 2012. Interchanging Lexical Resources on the Semantic Web. Language Resources and Evaluation, 46(4):701-719.

McCrae, John P., Chiarcos, Christian, Bond, Francis, Cimiano, Philipp, Declerck, Thierry, Melo, Gerard de, Gracia, Jorge, Hellmann, Sebastian, Klimek, Bettina, Moran, Steven, Osenova, Petya, Pareja-Lora, Antonio, Pool, Jonathan. 2016. The Open Linguistics Working Group :: Developing the Linguistic Linked Open Data Cloud. Proceedings of the

Tenth International Conference on Language Resources and Evaluation (LREC'16).

McCrae, John P. Tiberius, Carole, Khan, Anas Fahad, Kernerman, Ilan, Declerck, Thierry, Krek, Simon, Monachini, Monica, Ahmadi, Sina. 2019. The ELEXIS Interface for Interoperable Lexical Resources. In Proceedings of the eLex 2019 conference. Biennial Conference on Electronic Lexicography (eLex-2019) Electronic lexicography in the 21st century. October 1-3 Sintra Portugal Pages 642-659 ISBN 2533-5626 Lexical Computing CZ, s.r.o Brno 10/2019.

Melby, Alan K. 2015. TBX : A terminological exchange format for the translation and localisation industry. In Kockaert, Hendrik J. and Frieda Steurs (eds.). Handbook of Terminology. Volume 1. Amsterdam, Philadelphy : John Benjamins Publishing.

Rodiguez-Doncel,Victor, Casanovas, Pompeu. 2018. A Linked Data Terminology for Copy-right Based on Ontolex-Lemon : AICOL. International Workshops 2015-2017. 410-423.

Romary, Laurent. 2015. TEI and LMF crosswalks. JLCL (30)

Romary, Laurent. 2014. "TBX goes TEI- Implementing a {TBX} basic extension for the Text Encoding Initiative guidelines", In *CoRR*. Accessed September 19. http://arxiv.org/abs/1403.0052

Terminorgs. 2014. TBX-BASIC. Version 3.1, Terminology for Large Organizations. Accessed September 19 http://www.terminorgs.net/downloads/TBX_Basic_Version_3.1.pdf

Tiberius, Carole and Declerck, Thierry. 2017. A lemon Model for the ANW Dictionary. In Proceedings of the eLex 2017 conference, Pages 237-251, Leiden, Netherlands, Lexical Computing CZ s.r.o., INT, Trojína and Lexical Computing, Brno, Czech Republic.

Résumé

Dans cette contribution, nous décrivons la (ré)utilisation de l'infrastructure de recherche ELEXIS pour la lexicographie afin de traiter des données terminologiques. Nous présentons les aspects centraux de l'infrastructure ELEXIS et les normes qu'elle applique et développe. Nous présentons également TBX, qui est la principale norme utilisée pour représenter les données terminologiques. Nous décrivons en détail les spécifications d'OntoLex-Lemon, qui résultent d'un "Community Group" du W3C et qui jouent un rôle central dans notre travail, notamment en décrivant les données terminologiques au sein d'une infrastructure pour la lexicographie, car ces spécifications permet-

tent de relier les systèmes d'organisation des connaissances à une description lexicale détaillée. Pour illustrer cette capacité, nous utilisons des données de terminologie multilingues, à l'origine codées en TBX, provenant du domaine de la gestion des risques.

A good TACTIC for lexicographical work: football terms encoded in TEI Lex-0

Ana Salgado, Rute Costa

NOVA CLUNL, Centro de Linguística da Universidade NOVA de Lisboa
Avenida de Berna, 26-C 1069-061 Lisboa, Portugal
anasalgado@campus.fcsh.unl.pt
rute.costa@fcsh.unl.pt
https://clunl.fcsh.unl.pt/equipa/ana-salgado/
https://clunl.fcsh.unl.pt/equipa/rute-costa/

Abstract. Terms are a significant part of lexicographical nomenclatures in general language dictionaries. In this paper, we focus on how football terms are treated in three Academy Dictionaries – Portuguese, French, and Spanish – and draw some conclusions about the lexicographical decisions taken in the three languages. After identifying every position football players can have on the field, we verify whether the dictionaries above include these terms. We propose the TEI encoding of the term "defesa" (defence), which designates a position occupied by football players on the field. Bearing in mind concepts such as reusability and interoperability, we intend to present: 1) a comparison of football terms in the three dictionaries; 2) TEI Lex-0 dictionary encoding, a streamlined standard to facilitate interoperability; 3) a consistent TEI modelling and description of the microstructural elements of lexicographical entries. In the end, we draw some conclusions.

1. Introduction

Lexicography is conceived as a field that deals mainly with lexical units (words) but also with specialised lexical units (terms). While Lexicography and Terminology are two different scientific disciplines – with different theoretical-epistemological backgrounds – they have in common the fact that they deal with terms, albeit with different aims more often than not. This means

that working in Terminology and Lexicography requires different approaches since the social, cultural or economic purposes are not the same.

Lexicographers follow mostly a semasiological perspective (from words to senses), and terminologists, mostly concept-oriented, combine conceptual organisation and linguistic analysis where the definition of the concept is central with the view to reduce linguistic ambiguities. In 2013, Costa stated that Terminology and Lexicography should be seen as complementary regarding the methods they use. Bowker (2018, p. 149), arguing for the relation between these fields, sees advantages in the fact that "lexicographers and terminologists continue to work together to tackle new challenges and embrace new opportunities".

Getting to know the domain and subsequently organising it are two requisite activities for a rapid and systematic identification of the basic concepts, which will result in a better description of the terminology. Bearing in mind that we are working with a specialised domain, the intervention of the expert is necessary to aid in the task of organising knowledge and to validate the descriptions and definitions of terms (Silva & Costa, 2019). This facilitates a more accurate encoding by allowing a tidier classification of the data depending on each element.

The primary purpose of this paper is to show how football terms are treated in the three Academy dictionaries that comprise our lexicographical corpus (i.e. Portuguese, French, and Spanish languages)[1] (section 2). After identifying every position football players can have on the field, we confirmed whether the Academy dictionaries include these football terms (section 3). A contrastive study of the corpus allowed us to observe the data and draw conclusions about the lexicographical work performed in the three languages, gaining some insights into the lexicographical decisions that have been made in the three different dictionaries.[2] Subsequently, we applied and tested the latest version of the TEI Lex-0 guidelines by representing the term "defesa" (defence), which designates a position occupied by players on the field (section 4). Aiming at providing guidelines for a structurally organised and consist-

1 DLPC = *Dicionário da Língua Portuguesa Contemporânea*, 2001, Academia das Ciências de Lisboa; DAF = *Dictionnaire de l'Académie Française*, 2019, Académie Française, http://www.dictionnaire-academie.fr/; DLE = *Diccionario de la Lengua Española*, 2019, Real Academia Española, www.rae.es/rae.

2 This research is part of the PhD project of the first author that intends to design and test a set of methodological guidelines that can facilitate the inclusion of terms in a general language dictionary by systematising information regarding a single domain.

ent processing of lexicographical information, and bearing in mind concepts such as reusability and interoperability, we intend to present: 1) arguments in favour of using TEI Lex-0 dictionary encoding, a streamlined standard to facilitate interoperability; and 2) a consistent encoding and description of the microstructural elements of lexicographical entries, exemplifying the representation of lexicographical content using the term "defesa" (defence).

2. The lexicographical corpus

Recognising the importance of national academies that aim to create a dictionary in order to preserve the language, we decided to build a lexicographical corpus consisting of Academy works. The Academies of Sciences have undertaken dictionaries that are considered official authorities on the usages and the vocabulary of a language.

Our lexicographical corpus is comprised of three dictionaries published by different Academies: the *Dicionário da Língua Portuguesa Contemporânea* by the Academia das Ciências de Lisboa (DLPC), the *Diccionario de la Lengua Española* by the Real Academia Española (DLE), and the *Dictionnaire de l'Académie Française* by the Académie Française (DAF). These are general language contemporary dictionaries with printed editions and a descriptive nature with a normative concern addressed to a vast audience. The reason why we decided to create a contrastive corpus is justified by the fact that although the languages are different and the dictionaries themselves are also different, they share similar problems.

The content of these dictionaries is written in the languages of origin of each institution, i.e. Portuguese, Spanish, and French. These three dictionaries are available online[3], accessible for free, and are updated continuously.

In Portugal, despite the successive attempts of the Academy, only in 2001, under the coordination of Malaca Casteleiro, did the Academia das Ciências de Lisboa publish a complete dictionary (from A to Z) for the first time in a two-volume paper version: *Dicionário da Língua Portuguesa Contemporânea* (DLPC). A new digital version – a task that has been undertaken by a team working in Natural Language Processing (NLP) at the University of Minho

3 The DLPC is currently only available in-house. However, as the coordinator of the new dictionary of the Academia das Ciências de Lisboa, the first author is making steps towards its public availability.

– is the basis of the ongoing review of this dictionary, which now counts with the participation of NOVA CLUNL.[4]

The *Diccionario de la Lengua Española* is the widest normative dictionary of Spanish, published and created by the RAE. The most recent edition is its 24th, and it has been online since 2005.

The official dictionary of the French language, known as the *Dictionnaire de l'Académie Française* (DAF), served as a model for other dictionaries for many years, and the new version was made available on 7 February 2019, integrating the 9th edition in progress, from letters A to S.

3. Football domain

Football is the domain we have chosen to test the proposal for a set of future methodological guidelines for the lexicographical processing of terms. Our interest in football arises from the fact that it has been the most popular sport on the planet since the end of the 19th century; a sport with worldwide expansion via different societies in every continent. It is estimated that 250 million people are directly involved in football and that 1.4 billion people in the world have some interest in football (Morris, 1985).

We also aim to observe popularisation of the terms, namely the transition of a term into the vocabulary of everyday language. Many football terms have been adopted in everyday language, such as "canto" (corner) or "defesa" (defence) for instance.

This sport is also often referred to as 11-player football because it is played between two teams of 11 players each as seen on each of the definitions of the term in the three dictionaries that comprise our corpus (Fig. 1): "onze jogadores" (DLPC), "onze joueurs" (DAF), "once jugadores" (DLE):

4 The Natural Language Processing group of the Computer Science Department of the University of Minho has been developing the technological support of the new digital edition of the DACL, counting on the participation of Alberto Simões from IPCA (Instituto Politécnico do Cávado e do Ave), responsible for the technological support, José João Almeida, and the consultancy of Álvaro Iriarte Sanromán, both from the University of Minho. The participation of NOVA CLUNL (Linguistic Research Center of NOVA University of Lisbon) is related to its transition into the TEI LEX-0 format.

futebol [futibɔ́l]. *s. m.* (Do ingl. *football*). *Desp.* **1.** Modalidade desportiva em que jogam duas equipas, de onze jogadores cada uma, que procuram controlar e introduzir a bola na baliza da equipa adversária. *O futebol é a modalidade desportiva mais popularizada em Portugal. Jogo, desafio de +; clube, equipa, jogador de +; campo de +; jogar +.*

fútbol

Tb. **futbol.**

Del ingl. *football.*

1. m. Juego entre dos equipos de once jugadores cada uno, cuyo objetivo es hacer entrar en la portería contraria un balón que no puede ser tocado con las manos ni con los brazos, salvo por el portero en su área de meta.

FOOTBALL (se prononce *foutbol*) **nom masculin**

XVIIIe siècle. Mot anglais composé de *foot*, « pied », et *ball*, « balle, ballon ».

Désignait initialement tout jeu de ballon opposant deux équipes. Aujourd'hui, est réservé au jeu mettant aux prises deux équipes de onze joueurs qui doivent faire entrer un ballon, sans le toucher de la main ou du bras, dans les buts défendus par le camp adverse. *Un terrain de football. Un joueur, un club de football. Football amateur, professionnel. La Coupe de France, la Coupe du monde de football.*

FIG. 1 – *Entry "futebol/football/fútbol" (DLPC, DAF, DLE)*

3.1. The positions of football players on the field

The 11 football players occupy specific positions on the field which are connected with specific terms (Fig. 2):

FIG. 2 – *Football players occupy different positions on the field*

For quick identification of all possible positions of football players on the field, we present Fig. 3:

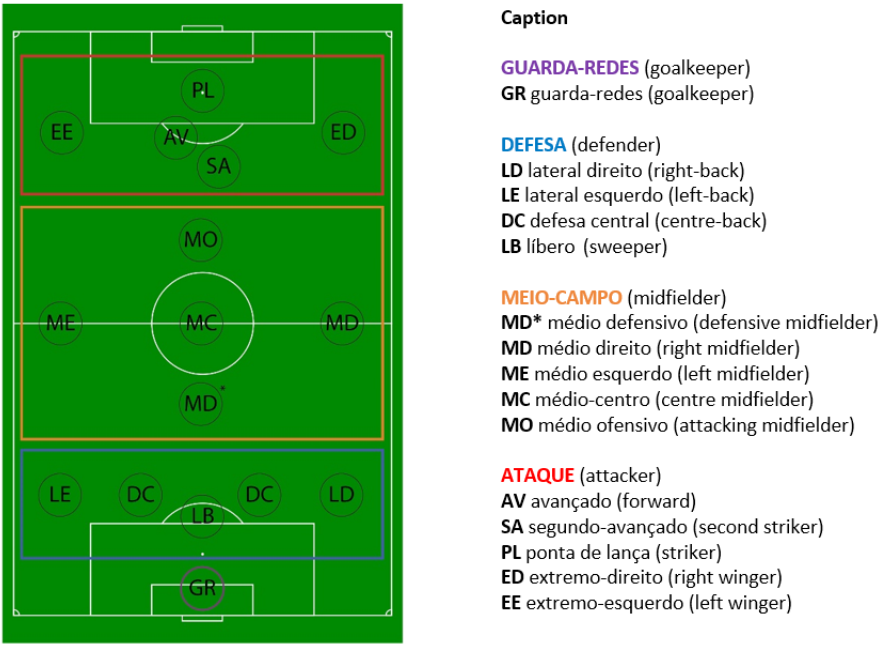


FIG. 3 – *Positions of football players on the field*

The positions of the players indicate the specific function that they play on the field; they are typically associated with the tactical scheme used, and can be divided into four fundamental positions: (1) goalkeeper (GR); (2) defender positions (LD, LE, DC, LB); (3) midfielder positions (MD*, MD, ME, MC, MO); (4) attacker positions (AV, SA, PL, ED, EE).

In Tab. 1, and retrieving Fig. 3, we have listed some terms in Portuguese related to positions with their equivalents in Spanish and French[5]. We have marked their presence (✓) or absence (-) in our lexicographical corpus.

5 The translation into English is used here only for the purpose of making this communication clearer.

Abbrev.	Portuguese	French	Spanish	English [literal translation]	DLPC	DAF	DLE
GUARDA-REDES (guardien de but, portero, goalkeeper)							
GR	guarda-redes	gardien de but	portero	goalkeeper	✓	✓	✓
DEFESA (defense, defensa, defender)							
LD	lateral direito	arrière latéral droit	lateral derecho	right-back	–	–	–
LE	lateral esquerdo	arrière latéral gauche	lateral izquierdo	left-back	–	–	–
DC	defesa central	défenseur central	defensa central	centre-back	–	–	–
LB	líbero	Libéro	líbero	sweeper	✓	–	✓
MEIO-CAMPO (centro del campo, milieu du terrain, midfield)							
MD*	médio defensivo	milieu défensif	volante de corte	defensive midfielder	–	–	–
MD	médio direito	milieu droit	volante externo derecho	right midfielder	–	–	–
ME	médio esquerdo	milieu gauche	volante externo izquierdo	wide midfielder	–	–	–
MC	médio-centro	milieu central	volante central	centre miedfielder	–	–	–
MO	médio ofensivo	milieu offensif	volante ofensivo	attacking midfielder	–	–	–
ATAQUE (ataque, attaque, attack)							
AV	avançado	avant	delantero	centre forward	✓	–	✓
SA	segundo-avançado	deuxième avant	segundo delantero	second striker	–	–	–
PL	ponta de lança	buteur	punta	striker	✓	–	✓
ED	extremo-direito	aillier droit	extremo derecho	right winger	–	–	–
EE	extremo-esquerdo	aillier gauche	extremo izquierdo	left winger	–	–	–

Tab. 1 – *Terms referring to positions occupied by football players on the field*

Looking at Tab. 1, we can see that only the term "goalkeeper" is recorded in all the dictionaries. Most terms that designate the positions of the players are not recorded in our dictionaries – e.g., "right-back", "left-back", "centre-back", "right winger", and "left winger". We may argue that this happens because we are dealing with polylexical units, such as "left back", and not just with monolexical units, such as "back" in English, "lateral" in Portuguese, "latéral" in French, and "lateral" in Spanish. In consequence, we decided to search for these units in our lexicographical corpus. The unit "lateral", when related to football, figures in the DLPC ("Fut. Jogador que actua junto da linha lateral do campo." [6]) and in the DLE ("Dicho de un futbolista o de un jugador

6 Player acting near the sideline.

de otros deportes: Que actúa junto a las bandas del terreno de juego con funciones generalmente defensivas."[7]), but is absent from the DAF.

To avoid such inconstancies, a terminological approach to the domain would be of major help. Building a concept system by identifying the relations between the concepts that embody the positions occupied by football players would allow the lexicographers to compile all the terms designating them. A conceptual approach to domains prevents lexicographers from missing essential terms of a terminology.

A term that is included in all these dictionaries, "goalkeeper", raises some controversial questions. Although the DLPC uses "Fut." (football) as a domain label listed in the abbreviation list, in the case of "guarda-redes", the domain label used is "Desp." (sports) ("Desp. Jogador que, no jogo do futebol, andebol, hóquei... ocupa o último posto de defesa, entre os postes da baliza, tentando impedir a marcação de golos"[8]). This happens because the 'definition' presented above is not only related to the football domain, but includes other sports. In the DLE, "portero" is not identified by any label ("Jugador que en algunos deportes defiende la portería de su bando"[9]). Finally, the DAF use the "Sports" label ("SPORTS. Gardien de but, joueur assurant la défense des buts dans certains jeux de ballon"[10]).

It seems clear that the DLPC and the DAF distance themselves from the DLE by using the domain label to differentiate meanings or contextualize them, merely specifying the domain of the meaning. In fact, any criterium can be validated as long as it is applied uniformly.

4. Text Encoding Initiative (TEI)

In order to make content interoperable and reusable, we decided to follow the Guidelines of the Text Encoding Initiative (TEI), which is a *de facto* standard in digital edition. These guidelines provide a dictionary module (TEI Consortium[11]) and have been used in numerous dictionary projects for

7 Said of a football player or a player of other sports: One that acts alongside the sidelines with generally defensive functions.
8 Player who, in football, handball, hockey..., occupies the last defense position between the goal posts, trying to prevent the scoring of goals.
9 Player who in some sports defends the goal of their side.
10 Goalkeeper, player defending goals in certain ball games.
11 https://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html

born-digital lexicographical data (Budin *et al*., 2012) or retrodigitised projects (Bohbot *et al*., 2018).

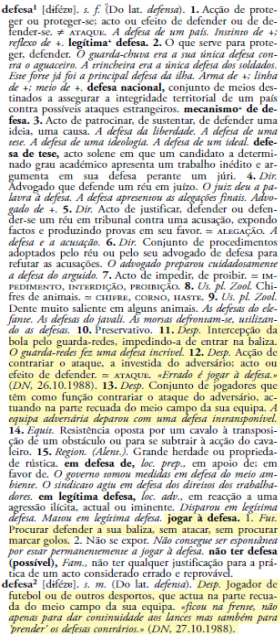
However, since the TEI Guidelines present numerous and flexible encoding possibilities, a new version specifically for dictionary encoding is now being discussed –TEI Lex-0[12] (Romary and Tasovac, 2018; Bański *et al*., 2017). We have chosen to test this target format, a streamlined standard to facilitate interoperability and some best-practice guidelines for NLP purposes.

TEI Lex-0 will not replace the "Dictionaries" chapter of the TEI Guidelines; instead it is being discussed as a target format that will standardise the existing heterogeneously encoded lexical resources and is being tested by numerous language dictionaries (Salgado *et al*., 2019b).

4.1. TEI Lex-0 encoding

The application of TEI Lex-0 will be demonstrated with samples of the term "defence" (Figures 4, 5, and 6) of the DLPC, the DAF, and the DLE. A consistent encoding and description of the microstructural components of the lexicographical article was the first step of this codification. Looking at the three dictionaries, the entries "defesa" (Fig. 4), "défense" (Fig. 5), and "defensa" (Fig. 6) include the same following elements: headword; etymology; part-of-speech (in the DLE this information is given in sense); a series of numbered meanings (polysemy). Depending on the criteria adopted in each of the dictionaries, we can also have domain labelling and other type of usage information, synonyms, collocations, and examples of usage.

12 TEI Lex-0 — A baseline encoding for lexicographical data: https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html



defesa[1] [difézɐ]. *s. f.* (Do lat. *defensa*). **1.** Acção de proteger ou proteger-se; acto ou efeito de defender ou de defender-se. ≠ ATAQUE. *A defesa de um país. Insistiu de +; reflexo de +.* **legítima* defesa. 2.** O que serve para proteger, defender. *O guarda-chuva era a sua única defesa contra o aguaceiro. A trincheira era a única defesa dos soldados. Esse forte já foi a principal defesa da ilha. Arma de +; linha de +; meio de +.* **defesa nacional,** conjunto de meios destinados a assegurar a integridade territorial de um país contra possíveis ataques estrangeiros. **mecanismo* de defesa. 3.** Acto de patrocinar, de sustentar, de defender uma ideia, uma causa. *A defesa da liberdade. A defesa de uma tese. A defesa de uma ideologia. A defesa de um ideal.* **defesa de tese,** acto solene em que um candidato a determinado grau académico apresenta um trabalho inédito e argumenta em sua defesa perante um júri. **4.** *Dir.* Advogado que defende um réu em juízo. *O juiz deu a palavra à defesa. A defesa apresentou as alegações finais. Advogado de +.* **5.** *Dir.* Acto de justificar, defender ou defender-se um réu em tribunal contra uma acusação, expondo factos e produzindo provas em seu favor. = ALEGAÇÃO. *A defesa e a acusação.* **6.** *Dir.* Conjunto de procedimentos adoptados pelo réu ou pelo seu advogado de defesa para refutar as acusações. *O advogado preparou cuidadosamente a defesa do arguido.* **7.** Acto de impedir, de proibir. = IMPEDIMENTO, INTERDIÇÃO, PROIBIÇÃO. **8.** *Us. pl. Zool.* Chifres de animais. = CHIFRE, CORNO, HASTE. **9.** *Us. pl. Zool.* Dente muito saliente em alguns animais. *As defesas do elefante. As defesas do javali. As morsas defrontam-se, utilizando as defesas.* **10.** Preservativo. **11.** *Desp.* Intercepção da bola pelo guarda-redes, impedindo-a de entrar na baliza. *O guarda-redes fez uma defesa incrível.* **12.** *Desp.* Acção de contrariar o ataque, a investida do adversário; acto ou efeito de defender. = ATAQUE. *«Errado é jogar à defesa.» (DN,* 26.10.1988). **13.** *Desp.* Conjunto de jogadores que têm como função contrariar o ataque do adversário, actuando na parte recuada do meio campo da sua equipa. *A equipa adversária deparou com uma defesa intransponível.* **14.** *Equit.* Resistência oposta por um cavalo à transposição de um obstáculo ou para se subtrair à acção do cavaleiro. **15.** *Region. (Alent.).* Grande herdade ou propriedade rústica. **em defesa de,** *loc. prep.*, em apoio de; em favor de. *O governo tomou medidas em defesa do meio ambiente. O sindicato agiu em defesa dos direitos dos trabalhadores.* **em legítima defesa,** *loc. adv.*, em reacção a uma agressão ilícita, actual ou iminente. *Disparou em legítima defesa. Matou em legítima defesa.* **jogar à defesa.** 1. *Fut.* Procurar defender a sua baliza, sem atacar, sem procurar marcar golos. 2. Não se expor. *Não consegue ser espontânea por estar permanentemente a jogar à defesa.* **não ter defesa (possível),** *Fam.*, não ter qualquer justificação para a prática de um acto considerado errado e reprovável.

defesa[2] [difézɐ]. *s. m.* (Do lat. *defensa*). *Desp.* Jogador de futebol ou de outros desportos, que actua na parte recuada do meio campo da sua equipa. *«ficou na frente, não apenas para dar continuidade aos lances mas também para 'prender' os defesas contrários.» (DN,* 27.10.1988).

FIG. 4 – *Entry "defesa" (DLPC)*

I. **DÉFENSE** nom féminin

XII^e siècle. XIX^e siècle, comme terme de droit, au sens de « la partie qui se défend ». Emprunté du bas latin *defensa*, « défense », participe passé féminin substantivé de *defendere*, « défendre ».

1. Action de défendre quelqu'un ou de se défendre contre une attaque. *Venir, courir à la défense* [...]

3. Par analogie. MÉDECINE. *Défense musculaire*, contraction douloureuse des muscles (s'emploie surtout à propos de la paroi abdominale, où la défense musculaire témoigne le plus souvent d'une inflammation péritonéale). – PHYSIOLOGIE. *Les défenses de l'organisme*, ses moyens de résistance à l'invasion microbienne, virale, parasitaire, etc. *L'immunologie étudie le système de défense de l'organisme.* – PSYCHOLOGIE. *Une réaction de défense. L'instinct de défense. Mécanismes de défense*, moyens par lesquels le sujet cherche inconsciemment à se protéger. – JEUX DE BALLON. Action de défendre les buts d'une équipe, de résister aux attaques de l'équipe adverse et, par extension, en parlant d'un sportif, action de s'opposer aux offensives de l'adversaire. Par métonymie. *La défense*, les lignes arrière d'une équipe, chargées de défendre les buts. – ÉQUITATION. *Les défenses d'un cheval*, les mouvements par lesquels un cheval se dérobe à l'action de son cavalier. – MARINE. *Défenses*, morceaux de bois, tampons de cordage, pneus usagés, cylindres de matière plastique disposés le long du bord d'un navire afin d'amortir les chocs ou d'empêcher les frottements (on dit aussi *Pare-battage*).

FIG. 5 – *Entry "défense" (DAF)*

defensa

Del lat. tardío *defensa.*

1. f. Acción y efecto de defender o defenderse.
2. f. Arma, instrumento u otra cosa con que alguien se defiende en un peligro.
3. f. Amparo, protección, socorro.
4. f. Obra de fortificación que sirve para defender una plaza, un campamento, etc. U. m. en pl.
5. f. Jugada del tresillo en la que un jugador sustituye en sus derechos y deberes al hombre que rinde la jugada.
6. f. Mecanismo natural por el que un organismo se protege de agresiones externas. U. t. en pl. con el mismo significado que en sing.
7. f. **línea defensiva.**
8. f. En ajedrez, conjunto de jugadas previstas que constituyen una determinada estrategia.
9. f. Razón o motivo que se alega en juicio para contradecir o desvirtuar la pretensión del demandante.
10. f. Abogado defensor del litigante o del reo. U. m. en pl. cuando hay varios reos en el mismo juicio.
11. f. *Cuba, Méx., R. Dom.* y *Ur.* **parachoques.**
12. f. pl. Colmillos del elefante, cuernos del toro, etc.
13. f. pl. *Mar.* Conjunto de objetos que se cuelgan del costado de la embarcación para que esta no roce o golpee contra el muelle u otra embarcación.
14. m. y f. Cada uno de los jugadores que forman la línea defensiva.

FIG. 6 – *Entry "defensa" (DLE)*

Focusing on the form element that contains the orthographic, phonetic, and grammatical information of the entry (headword), all the elements of the entries are distinguished as clearly as possible:

```
<entry xml:lang="pt"
xml:id="defesa_1"
n="1">
<form type="lemma">
<orth>defesa</orth>
<pron>dɨfézɐ</pron>
</form>
<gramGrp>
<gram type="pos"
norm="NOUN">s.</
gram>
<gram
type="gen">f.</gram>
</gramGrp>
<etym>
[…]
</etym>
[…]
</entry>
```

```
<entry xml:lang="fr"
xml:id="défense">
<form type="lemma">
<orth>défense</orth>
</form>
<gramGrp>
<gram type="-
pos" norm="-
NOUN">nom</gram>
<gram
type="gen">féminin</
gram>
</gramGrp>
<etym>
[…]
</etym>
[…]
</entry>
```

```
<entry xml:lang="es"
xml:id="defensa">
<form type="lemma">
<orth>defensa</orth>
</form>
<etym>
[…]
</etym>
[…]
</entry>
```

Fig. 7 – *Entry "defesa/défense/defensa" (DLPC, DAF, DLE) encoded in TEI Lex-0*

From these encoding examples, we highlight that the TEI Lex-0 schema only uses entry, the basic element of the dictionary's microstructure, once we have constrained the general structure of a lexical entry – in our schema, entryFree, superEntry and re (related entry) of the current Guidelines are not used. In Fig. 7, entry contains the following elements – form, etym, and gramGrp – and requires the attributes @xml:id and @xml:lang, the appropriate language code (BCP 47) – "pt" (DLPC,; "fr" (DAF), or "es" (DLE). Dictionary entries always start with a lemma (headword), a canonical form. The lemma form is encoded as form @type="lemma".

To specify the morpho-syntactic properties of the entry, we use gramGrp, which includes the part-of-speech of the entry (pos) and further specifications, such as the morphological gender (gen) of a lexical item. In

the example, for part-of-speech we use the universal dependencies tagset.[13] In addition, we highlight the fact that the DLPC has also phonetic transcription (pron) and a numerical index (number), and, in the DLE, the part of speech (a noun) is inferred from the attribution of the female gender and is marked in sense.

13 In the ELEXIS context, a @norm attribute is mandatory to specify a normalised (UD) part of speech value. For more details: https://universaldependencies.org/u/pos/

```
<sense xml:id="defesa_11" n="11">
<usg type="domain">-Desp.</usg>
<def>Intercepção da bola pelo guarda-redes, impedindo-a de entrar na baliza.</def>
<cit type="example">
<quote>O guarda-redes fez uma defesa incrível.</quote>
</cit>
</sense>
<sense xml:id="defesa_12" n="12">
<usg type="domain">Desp.</usg>
<def>Acção de contrariar o ataque, a investida do adversário; acto ou efeito de defender.</def>
<xr type="antonymy">
<ref type="sense">ataque</ref>
</xr>
<cit type="example">
<quote>«Errado é jogar à defesa.»</quote>
<bibl> <title>DN</title>
<date>26.10.1998</date></bibl>
</cit>
</sense>
[…]
</entry>
```

```
<sense xml:id=»défense_3.4» n=»3»>
<usg type=»domain»>-JEUX DE BALLON.</usg>
<def>Action de défendre les buts d'une équipe, de résister aux attaques de l'équipe adverse et, par extension, en parlant d'un sportif, action de s'opposer aux offensives de l'adversaire.</def>
</sense>
<sense xml:id="defensa_3.5" n="3">
<usg type=»meaningType»>Par métonymie.</usg>
<cit type=»example»>
<quote>La défense</quote>
<title>,</title>
<def>les lignes arrière d'une équipe, chargées de défendre les buts.</def>
</sense>
[…]
</entry>
```

```
<sense xml:id="defensa_7" n="7">
<gramGrp>
<gram type="gen">f.</gram>
</gramGrp>
línea defensiva.
</sense>
[…]
<sense xml:id="defensa_14" n="14">
<gramGrp>
<gram type="gen">m.</gram>
<lbl>y</lbl>
<gram type="gen">f.</gram>
</gramGrp>
<def>Cada uno de los jugadores que forman la línea defensiva.</def>
</sense>
[…]
</entry>
```

FIG. 8 – *Entry "defesa/défense/defensa" (DLPC, DAF, DLE) encoded in TEI Lex-0*

When dealing with football terms, the associated lexical items belong to a specific domain. This usage information is usually identified in dictionaries through domain labelling (Salgado and Costa, 2019; Salgado *et al*., 2019a). In Fig. 4, senses 11, 12, and 13 are marked with "Desp." (the abbreviated form of "Desporto" [Sports]). In Fig. 5, the sense related to football is marked with the expanded form "JEUX DE BALLON". Lastly, in Fig. 6, senses number 7 and 14, which refer to football, are not annotated. In TEI-style encoding, domain is a marker that identifies the specialised field of a lexical item – in TEI Lex-0 we use the element usg to mark usage information such as this (Fig. 8). Looking at other examples from Tab. 1, you can detect inconsistencies, since, for example, in the DLPC, both sports and football domains are used (e.g. "defesa-direito" is marked with the football domain, while "ponta-direita", a synonym of "extremo-direito", again a position occupied by a football player on the field, is marked as sports). Finally, we will also highlight multiword expressions, such as "jogar à defesa" (to play defence), which are treated as entries. We argue that this information should be standardised to a single value.

5. Concluding remarks

We want to conclude with four final remarks:

1. Our research has strictly lexicographical purposes, using Terminology's methodologies as a contribution to the definition of guidelines and a methodology for the selection, inclusion, and processing of terms in general language dictionaries, namely the Academy Dictionaries we have analysed here, proposing a new dictionary model that, in a harmonised and balanced way, combines lexicographical methodologies and terminological methodologies.
2. Combining conceptual organisation and linguistic analysis is a necessity. Getting to know the domain and subsequently organising it are two requisite activities for a rapid and systematic identification of the basic concepts, which will result in a better description of the lexicon. This facilitates encoding by allowing a tidier classification of the data depending on each element, such as entry or sense, or @type "domain"> attribute. Bearing in mind that we are working with specialised domains, the intervention of the expert is necessary to aid in the task of organising knowledge, which will result in more accurate encoding.
3. By examining the codification of the term "defesa", we have confirmed that the Lex-0 TEI meets our research needs. After encoding the

microstructure of the dictionary, we can guarantee the interoperability and reusability of the information. The advantage of applying TEI Lex-0 lies in the fact that lexicographers and terminologists alike are currently making efforts to apply TEI to the ongoing revision of ISO LMF (Romary, 2015). Given its (still) non-standard nature, it can be changed to accommodate relevant dictionary structures. We intend to demonstrate that the results obtained are useful for computational lexical encoding and may serve the purpose of NLP.

4. Finding that the dictionaries that make up our lexicographical corpus share common problems regarding the example we worked with ("defesa") has led us to suggest that it would be interesting to present identical solutions for all of them. The reason why we decided to create a contrastive corpus is justified by the fact that although the languages are different and the dictionaries themselves are also different, they share similar problems. A contrastive analysis of the three dictionaries allowed us to confirm that. The solutions we have presented for the Portuguese example (DLPC) would be replicable in dictionaries of other languages since TEI is a recommendation for standardisation. An agreement between academies and other institutions would be desirable to systematise and optimise a new type of Lexicography that can provide a better representation of the entire European lexicographical heritage.

Acknowledgements

Research financed by Portuguese National Funding through the FCT – Fundação para a Ciência e Tecnologia as part of the project Centro de Linguística da Universidade NOVA de Lisboa – UID/LIN/03213/2020, and by the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015 (ELEXIS).

References

Bański, P., Bowers, J., and Erjavec, T. 2017. "TEI Lex-0 guidelines for the encoding of dictionary information on written and spoken forms". In *Electronic Lexicography in the 21st Century: Proceedings of eLex 2017 Conference*, 485-94.

Bohbot, H., Frontini, Luxardo, F. G., Khemakhem, M., and Romary, L. (2018). Presenting the Nénufar Project: a Diachronic Digital Edition of the Petit

Larousse Illustré. In *GLOBALEX 2018* – Globalex workshop at LREC2018, May 2018, Miyazaki, Japan, 1-6. <hal-01728328>.

Bowker, L. 2018. "Lexicography and terminology". In The Routledge Handbook of Lexicography, edited by P. Fuertes-Olivera, P. Abingdon: Routledge Handbooks Online.

Budin, G., Majewski, S. and Morth, K. 2012. "Creating Lexical Resources in TEI P5: A Schema for Multi-purpose Digital Dictionaries". In *Journal of the Text Encoding Initiative 3*, http://jtei.revues.org/522.

Costa, R. 2013. "Terminology and specialised lexicography: two complementary domains". In *Lexicographica*, 29, 29-42.

ISO 1087-1. 2000. Terminology work – Vocabulary – Part 1: Theory and application. Geneva: International Organization for Standardization.

Morris, D. 1985. *A Tribo do Futebol*. Lisboa: Publicações Europa-América.

Romary, L. 2015. "TEI and LMF crosswalks". In *Digital Humanities: Wissenschaft vom Verstehen*. Berlin: Humboldt Universität zu Berlin, <hal-00762664v2>.

Romary, L., Tasovac, T. 2018. "TEI Lex-0: A Target Format for TEI-Encoded Dictionaries and Lexical Resources". In Proceedings of the 8th Conference of Japanese Association for Digital Humanities, 274-275, https://tei2018.dhii.asia/AbstractsBook_TEI_0907.pdf.

Salgado, A., and Costa, R. 2019. Marcas temáticas en los diccionarios académicos ibéricos: estudio comparativo. *RILEX. Revista Sobre Investigaciones léxicas*, 2(2), 37-63. https://doi.org/10.17561/rilex.v2.n2.2.

Salgado, A. Costa, R. and Tasovac, T. 2019a. Improving the consistency of usage labelling in dictionaries with TEI Lex-0. In Lexicography: Journal of ASIALEX 6 (2), 133-156. DOI: https://doi.org/10.1007/s40607-019-00061-x.

Salgado, A., Costa, R., Tasovac, T. and Simões, A. 2019b. TEI Lex-0 In Action: Improving the Encoding of the Dictionary of the Academia das Ciências de Lisboa. In I. Kosem *et al*. (eds.), Electronic lexicography in the 21st century. Proceedings of the Electronic lexicography in the 21st century (eLex 2019): Smart lexicography, 417-433), 1-3 October 2019, Sintra, Portugal. Brno: Lexical Computing CZ, s.r.o. Retrieved from: https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_23.pdf.

Silva, R. and Costa, R. 2019. Accéder aux connaissances des experts par l'entremise de la médiation en Terminologie. In M. De Gioia e M. Marcon, eds., *L'essentiel de la médiation. Le regard des sciences humaines et sociales*. Peter Lang, 105-121. DOI: https://doi.org/10.3726/b16164. ISBN 978-2-8076-1088-0.

Tarp, S. 2008. *Lexicography in the Borderland between Knowledge and Non-Knowledge*. Tübingen: Niemeyer.

Dictionaries

Dicionário da Língua Portuguesa Contemporânea, 2001, João Malaca Casteleiro (coord.), 2 vols. Lisboa: Academia das Ciências de Lisboa & Editorial Verbo.

Diccionario de la Lengua Española (24.ª ed.). Real Academia Española [em linha], 2001-2020, www.rae.es/rae.

Dictionnaire de l'Académie Française (9.ª ed.). Académie Française, 2020, http://www.dictionnaire-academie.fr/.

Résumé

La complémentarité entre lexicographie et terminologie est le point de départ de cet article. Travaillant dans le domaine du football, cet article examine les termes du football dans trois dictionnaires de l'Académies de Sciences — portugais, français et espagnol — et présente l'encodage TEI du terme « defesa » (défense), qui désigne une position occupée par les joueurs sur le terrain. Après avoir identifié toutes les positions que les joueurs de football peuvent occuper sur le terrain, nous avons vérifié si les dictionnaires incluent ces termes. Une étude contrastive du corpus nous a permis d'observer les données et de tirer des conclusions sur le travail lexicographique réalisé dans les trois langues. En gardant à l'esprit des concepts tels que la réutilisabilité et l'interopérabilité, nous présentons : 1) une comparaison des termes dans les trois dictionnaires en analyse ; 2) le codage du dictionnaire TEI Lex-0, une norme *de facto* pour faciliter l'interopérabilité ; 3) une modélisation TEI cohérente et une description des éléments microstructuraux des entrées lexicographiques.

Protocole de construction d'un dictionnaire des médicaments pour les études en pharmacologie

François-Élie Calvier*, Bissan Audeh**,
Florelle Bellet***, Cédric Bousquet***

Service de Santé Publique et d'Information Médicale, Centre Hospitalier Universitaire de Saint-Étienne-Hôpital Nord, Saint-Étienne, France
*francois.calvier@laposte.net
**Audeh@emse.fr
***prenom.nom@chu-st-etienne.fr

Résumé. L'identification de médicaments dans un corpus textuel possède un grand nombre d'applications dans différents domaines de la pharmacologie. Néanmoins, aucun dictionnaire des médicaments, nécessaire à cette identification, n'existe en français. Dans cet article, nous proposons une méthode de construction itérative d'un tel dictionnaire mettant en correspondance le vocabulaire patient et le vocabulaire spécialisé utilisé en pharmacologie.

1. Introduction

L'identification de médicaments dans un corpus textuel possède un grand nombre d'applications dans différents domaines de la pharmacologie. C'est le cas, entre autres, de la pharmacovigilance qui s'intéresse à la détection, l'analyse et la compréhension des effets indésirables des médicaments ou de la pharmaco-épidémiologie qui s'intéresse à l'évaluation de l'efficacité, du risque, du bénéfice et de l'usage des médicaments. Depuis quelques années, les forums de discussion sont envisagés, dans le domaine de la pharmacovigilance, comme une source complémentaire aux déclarations spontanées des professionnels de santé et des patients. La détection des médicaments dans les messages extraits des forums fait partie des difficultés techniques associées à cette analyse.

En effet, l'identification de médicaments dans un texte est un processus en deux étapes : 1) repérer chaque mention d'un médicament et 2) associer

chaque mention au médicament auquel elle correspond. L'analyse pharmacologique des textes du corpus est alors facilitée (indexation, mise en évidence des mentions,…). Pour automatiser cette dernière étape, il est nécessaire de disposer d'un dictionnaire mettant en relation les éléments du corpus repérés comme mention de médicament et les entrées d'un thésaurus contenant l'ensemble des médicaments d'intérêt**.**

Dans cet article, nous nous attachons à la deuxième étape et proposons une méthode semi-automatique et itérative pour la construction d'un dictionnaire à partir d'une base de données médicamenteuse. Nous présentons, en section 2, notre contexte de travail et un état de l'art des travaux connexes à notre approche. Nous nous attachons ensuite, en section 3, à décrire notre approche puis, en section 4, les expérimentations que nous avons réalisées. Enfin, nous concluons sur une application concrète de cette approche dans le cadre du projet Vigi4MED (Vigilance dans les forums médicaments) et nous proposons plusieurs perspectives.

2. État de l'art

Les forums de discussion sur Internet sont un lieu d'échange dans lequel certains patients peuvent décrire leur traitement médicamenteux ou poser des questions aux autres participants. Comme de nombreux documents rédigés par les professionnels de santé (comptes rendus médicaux, déclarations spontanées d'effets indésirables liés à un médicament), il s'agit d'une ressource en texte libre insuffisamment exploitée en raison des difficultés liées à l'extraction, l'analyse et l'interprétation de ces données. Pour ce qui concerne les forums de discussion en langue française plusieurs études ciblées sur certaines classes médicamenteuses menées dans le domaine de la pharmacovigilance ont nécessité de parcourir manuellement les messages des patients pour rechercher des effets indésirables (Abou Taam, *et al*. (2014) Palosse-Cantaloube, *et al*. (2014) Lardon, *et al*. (2015), Golder, *et al*. (2015), Bagheri, *et al*. (2016)). La grande quantité de données disponibles dans ces forums rend impossible l'analyse manuelle de tous les messages. Par ailleurs, l'extraction automatique de connaissances médicales, dans ce cas nécessite la prise en compte des défis liés au traitement de textes libres et la gestion des corpus de grande taille. De plus, dans le domaine médical, il est nécessaire de considérer la coexistence de deux vocabulaires : un vocabulaire métier utilisé par les professionnels de santé et un vocabulaire patient utilisé par les usagers (Tapi Nzali, *et al*. (2015)). La réconciliation des deux vocabulaires nécessite

l'exploitation de dictionnaires définissant ces deux terminologies (Eholié *et al*. (2016)).

2.1. Vocabulaire patient

Il n'existe pas, à notre connaissance, de ressource en français décrivant les médicaments dans le vocabulaire patient. Lors de l'association de mentions de médicaments exprimés dans les forums à l'aide de ce vocabulaire patient avec le vocabulaire métier, des traitements de langage naturel comme la tokenization, la racinisation et l'étiquetage morphosyntaxique peuvent être appliqués pour améliorer l'alignement avec le dictionnaire. L'exploitation de corpus composés de textes libres a pour conséquence la possibilité de variations orthographiques. La complexité orthographique des termes du domaine médical renforce ce problème. Plusieurs approches ont été proposées dans le cadre de l'identification de médicaments pour traiter les variations orthographiques. Par exemple, les erreurs d'orthographe sur les noms de médicaments peuvent être caractérisées (Senger, *et al*. 2010). D'autre part, les noms de médicaments peuvent être transcrits phonétiquement avant comparaison (Levin, *et al*. 2007) (Pereira, *et al*. 2012) (Pimpalkute, *et al*. 2014). Enfin, Google spell checker permet de proposer des corrections orthographiques basées sur les erreurs fréquemment commises lors de requêtes sur le moteur de recherche. En exploitant cet outil, il est possible d'identifier les variations orthographiques fréquemment utilisées dans les requêtes Google (Zhou, *et al*. 2012). Une autre difficulté concerne l'existence de termes ambigus (Sirohi et Peissig 2005) (Hamon et Grabar 2010) pouvant appartenir à plusieurs champs sémantiques. Cette ambiguïté sémantique est généralement levée en traitant ces termes au sein du contexte dans lequel ils sont employés. Cependant, dans notre cadre de travail, les mentions de médicaments nous parviennent sans contexte pour l'étape d'association.

2.2. Vocabulaire métier

Au sein du vocabulaire métier, il existe plusieurs façons de désigner un médicament (Pahor, *et al*. 1994). Cependant, notre objectif est de proposer des résultats à destination d'experts en pharmacologie pour lesquels les médicaments sont exprimés en termes de principes actifs. Une mise en correspondance entre médicaments et principes actifs est donc nécessaire. Bien qu'il existe plusieurs bases de données métier comme la base de données publique

du médicament[1], les bases Claude Bernard[2], Thériaque[3] et Vidal[4], il n'existe pas, à notre connaissance, de terminologie de référence unifiée disponible pour les médicaments français, à l'image de RxNorm (Nelson, *et al*. 2011) aux États Unis. Ces bases de données ayant pour objectif l'aide à la prescription, elles doivent être adaptées pour notre cadre de travail.

3. Construction itérative du dictionnaire

Notre objectif est de mettre en relation un ensemble F de termes du vocabulaire utilisé dans les messages pour désigner des médicaments dans les forums avec l'ensemble C du vocabulaire métier constitué des principes actifs (par ex. paracétamol). C correspond à l'ensemble des termes qui intéressent les experts du domaine. Nous ne disposons pas de ressource définissant l'ensemble F mais nous disposons de bases de données mettant en relation les éléments de l'ensemble C avec le vocabulaire de l'ensemble M des noms de médicaments. Nous distinguons 3 sous-ensembles de F :

- La liste blanche contient tous les termes du vocabulaire patient pour lesquels au moins un principe actif est associé (par ex. baclo associé au baclofène ou doliprane associé au paracétamol).
- La liste grise contient tous les termes du vocabulaire patient pour lesquels il existe une ambigüité (par ex. jasmine). Comme pour les éléments de la liste blanche, au moins un principe actif est associé à chaque élément de la liste grise.
- La liste noire contient tous les termes du vocabulaire patient identifiés comme inexploitables parce qu'ils désignent une classe et pas un médicament (par ex. antidépresseur) ou ne correspondant pas à un médicament. Aucun principe actif n'est associé aux éléments de la liste noire.

L'ensemble F est composé d'éléments des ensembles C et M et d'éléments non référencés. Une phase d'initialisation consiste à insérer les éléments de M et de C dans F. Cette initialisation permet la mise en relation des éléments de F avec ceux de l'ensemble C et accélère ainsi la construction du dictionnaire. Cependant, cette initialisation n'est pas automatique car l'on souhaite détecter les éléments du vocabulaire métier qui sont ambigus (liste grise).

1 http://base-donnees-publique.medicaments.gouv.fr/
2 https://www.bcbdexther.fr/
3 http://www.theriaque.org/apps/contenu/accueil.php
4 https://www.vidal.fr/

À chaque itération, les mentions de médicaments sont comparées aux éléments de l'ensemble F. Pour chaque correspondance trouvée, un traitement par sous-ensemble de F est proposé. Les mentions ne trouvant pas de correspondance sont listées afin d'être étudiées puis ajoutées à l'ensemble F.

Dans notre cas, nous avons exploité, pour la phase d'initialisation, la base de données Thériaque (Husson 2008) du fait de sa disponibilité pour les auteurs.

3.2.1. Extraction des principes actifs et des médicaments

Dans la base Thériaque, un même principe actif peut être présent sous différentes formes. Par exemple, le paracétamol est présent sous forme de paracétamol, de paracétamol compap pvp3 et de paracétamol DC. Ces différentes entrées ne sont pas pertinentes pour la construction de notre dictionaire. En effet, d'une part, les patients les utilisent rarement. D'autre part les experts n'attendent pas autant de précision sur les médicaments extraits à partir de textes libres.

Dans un objectif de couverture optimale, tous les principes actifs de la base Thériaque sont ajoutés dans le dictionnaire, mais seuls les principes actifs de haut niveau pour lesquels aucune monographie n'est associée dans la base Thériaque (par exemple, paracétamol) sont ajoutés à la liste des termes cibles. Les principes actifs avec une monographie sont ajoutés à la liste blanche et liés avec les termes cibles auxquels ils correspondent.

Dans la base Thériaque, les médicaments sont présents sous deux formes :

- Les produits sont les noms des médicaments sans précision (par ex Efferalgan).
- Les spécialités sont les noms des médicaments avec toutes les précisions de forme, de dosage, de voie d'administration ou de population visée (par ex EFFERALGAN 500 mg VAN FR GRANULE SACH)

Les usagers utilisent davantage la dénomination des produits pharmaceutiques plutôt que celle des spécialités pharmaceutiques. Nous avons donc choisi de ne pas intégrer les spécialités. Par ailleurs, la base Thériaque étant à destination des hôpitaux, certains noms de médicaments ne sont pas exploitables pour l'identification de médicaments. Par exemple, les médicaments utilisés dans le cadre d'essais cliniques ou des préparations pharmaceutiques à usage interne sont désignés par des noms standardisés. Dans un objectif d'optimisation du dictionnaire, ces entrées ne sont pas ajoutées au diction-

naire. Leurs noms standardisés permettent une sélection rapide à partir d'expressions régulières.

3.2.2. Mise en correspondance entre médicaments et principes actifs

Cette mise en correspondance n'est pas triviale. En effet, il peut y avoir, pour un même produit pharmaceutique, différentes spécialités. Par exemple, le produit Efferalgan possède, entre autres, une spécialité Efferalgan van fr 250 mg granule sachet et une spécialité Efferalganodis 500 mg comprimé. Outre des variations de dosage et de forme, un même produit peut également posséder des spécialités ayant des compositions différentes en termes de principes actifs. Par exemple, il existe également une spécialité Efferalgan vitamine C comprimé effervescent. On peut alors considérer que les principes actifs associés à un produit correspondent à l'intersection des principes actifs présents dans toutes les spécialités de ce produit. Ce choix conduit à une sous-estimation du nombre d'occurrences des principes actifs présents uniquement dans certaines spécialités. Il est également possible, à l'inverse, de considérer que les principes actifs associés à un produit correspondent à l'union des principes actifs présents dans les spécialités de ce produit. Ce choix conduit à une sur estimation du nombre d'occurrences pour ces mêmes substances.

Le choix de mise en correspondance est un paramètre important pour la génération des résultats de l'exploitation de notre dictionnaire par l'utilisateur final. Cependant, il n'a pas d'influence sur le protocole que nous proposons ici.

3.1. Itération

Les syntagmes détectés dans un texte libre comme des traitements médicamenteux qui ne correspondent pas à une entrée du dictionnaire initial sont potentiellement des termes du vocabulaire patient que l'on cherche à aligner avec le dictionnaire. Nous illustrons, dans la Fig. 1, le processus d'identification des médicaments dans des textes libres et l'étape d'enrichissement du dictionnaire à l'aide de vocabulaire patient.

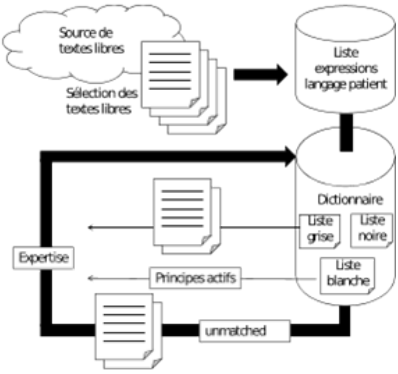


Fig. 1 Processus d'identification des médicaments

Chaque nouvelle extraction d'expressions de langage patient à partir de textes libre (par ajout de nouvelles sources ou par modification des algorithmes d'extraction) constitue une nouvelle itération. À chaque itération, le dictionnaire peut être enrichi. Les expressions du langage patient sont comparées aux entrées du dictionnaire constitué lors des itérations précédentes.

3.1.1. Normalisation des syntagmes

Dans les textes libres, il arrive que les noms de princeps soient suivis de caractères spéciaux tels que les Trade marks (®, ©, ™,…). Pour améliorer la mise en correspondance des syntagmes extraits avec les entrées du dictionnaire, il est nécessaire de traiter ces caractères spéciaux. Nous proposons, dans un premier temps, de les supprimer de la comparaison avec les entrées du dictionnaire.

De même, les caractères de séparation dans les noms de médicaments et de principes actifs (espace, tiret, blanc souligné) sont une source de nombreuses variations orthographiques. Nous avons choisi de traiter ce type de variations afin de limiter le nombre d'itérations nécessaires. Tous les caractères de séparation sont normalisés avant comparaison.

Une fois normalisées, les mentions de médicaments du corpus sont comparées aux entrées du dictionnaire.

3.1.2. Traitement des correspondances

Lorsqu'une correspondance est trouvée avec une entrée du dictionnaire, cette entrée appartient à l'une des trois listes qui constituent le dictionnaire. Lorsque la correspondance est trouvée avec un terme de la liste noire, la mention de médicament est automatiquement écartée. Lorsque la correspondance est trouvée avec un terme de la liste grise, le texte contenant la mention de médicament est proposé à l'expert pour validation. Lorsque la correspondance est trouvée avec un terme de la liste blanche, la mention de médicament est automatiquement validée. L'ensemble des correspondances validées sont ensuite amalgamées afin de produire une analyse statistique en termes de principes actifs pour l'expert.

3.1.3. Traitement des non-correspondances

Lorsqu'aucune correspondance n'est trouvée avec les entrées du dictionnaire, la mention de médicament détectée est alors potentiellement un terme du vocabulaire patient encore non répertorié. Il est nécessaire de présenter les textes contenant ces mentions à l'expert afin qu'il puisse indiquer comment ces mentions seront traitées lors des itérations suivantes. Plusieurs cas se présentent :

- Traitement des termes génériques : bien que pertinents du point de vue de la détection car il s'agit bien de mentions de médicaments, les termes génériques (tels que antibiotique, anti-inflammatoire, etc.) ne peuvent être rapprochés d'un élément de la liste cible. On les ajoute donc à la liste noire du dictionnaire.
- Traitement des abréviations et diminutifs : le domaine médical est sujet aux abréviations et acronymes du fait de la longueur et de la complexité des termes qu'il véhicule. Par ailleurs, les forums sont également vecteurs d'abréviations de tous ordres. Elles sont communes à des groupes d'usagers ou de professionnels de santé et nous faisons l'hypothèse que leur fréquence d'utilisation les rend identifiables. En plus de confirmer l'ajout dans la liste blanche de ce type de mentions de médicaments, l'expert doit préciser avec quel(s) terme(s) de la liste cible, elles sont associées. Par exemple, certains patients désignent le baclofène au moyen de l'abréviation baclo.
- Traitement des erreurs de détection : l'étape de repérage des mentions de médicaments peut parfois produire des résultats erronés. L'identification automatique de ces erreurs n'est pas triviale. De plus, l'aspect

automatique de l'étape de repérage peut conduire à une répétition d'une même erreur. L'ajout de ces erreurs en liste noire permet un traitement automatique d'une erreur de détection déjà rencontrée.

4. Expérimentations

Dans le cadre du projet Vigi4MED, nous avons travaillé à partir d'une source composée de plusieurs forums de discussion en langue française. Nous avons appliqué la chaîne de traitements décrite en Fig. 2 pour exploiter le potentiel d'un dictionnaire construit selon notre approche. Nous décrivons les données produites par les étapes de sélection des forums et d'extraction des syntagmes de langage patient, puis nous présentons les résultats que nous avons obtenus à partir de ces données.

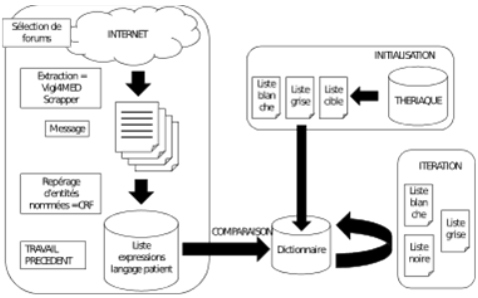


Fig. 2 Chaîne de traitements

4.1. Description des données

Pour notre étude, les messages ont été extraits entre 2000 et 2015 à la fois de forums médicaux (généralistes ou spécialisés) et de forums non médicaux dans lesquels des sujets médicaux sont abordés.

Les messages constituant les fils de discussions jugés pertinents après étude ont été extraits à l'aide de l'outil Vigi4MED Scrapper (Audeh, *et al.* 2017) développé par l'École Nationale Supérieure des Mines de Saint-Étienne.

Les mentions de médicaments dans ces messages sont identifiées à l'aide d'une méthode développée par le LIMSI (Laboratoire d'informatique pour la mécanique et les sciences de l'ingénieur) (Morlane-Hondère, Grouin et Zweigenbaum 2016).

Sur la période 2012 à 2014 commune à tous les forums, ces forums contiennent 1 241 565 messages pour lesquels au moins une mention de médicament a été détectée. En moyenne, chaque message contient environ 2 mentions de médicaments. Au total, 227 565 mentions de médicaments uniques ont été identifiées.

La période d'étude a été découpée pour réaliser 6 itérations à intervalles réguliers. Afin de tester l'apport de la phase d'initialisation, nous avons réalisé deux expérimentations sur ce même jeu de données : la première sans initialisation du dictionnaire, la seconde avec initialisation du dictionnaire à l'aide de Thériaque.

À chaque itération, nous évaluons le nombre de messages dans lesquels apparaît une mention de médicament. Nous sélectionnons les 100 mentions les plus fréquentes pour lesquelles aucune correspondance avec le dictionnaire n'a été trouvée. Ces mentions sont présentées à un expert pour classement dans les différentes catégories définies en section. Ces mentions sont ensuite intégrées au dictionnaire pour les itérations suivantes. L'objectif de ces expérimentations est de montrer l'intérêt de l'initialisation et de la construction par itérations du dictionnaire.

4.2. Résultats

Parmi les 227 565 mentions de médicaments du jeu de données, 3 021 correspondent à des médicaments ou des principes actifs décrits dans Thériaque. Au terme de chaque itération, 100 mentions sont étudiées. Parmi ces 600 mentions, 169 peuvent être validées automatiquement si le dictionnaire a été initialisé. L'expert a également identifié 16 mentions correspondant à du vocabulaire patient pour les médicaments, 53 à des classes de médicaments (« antibiotique », « contraceptif »,…) et 18 termes médicaux sans rapport avec les médicaments (« hôpital », « néonat »). Enfin 334 mentions résultent d'erreurs de détection.

La quasi-totalité (93 à 99 %) des mentions ajoutées au dictionnaire lors des itérations précédentes sont exploitées pour traiter automatiquement un nombre croissant de messages (de 36,24 % à la seconde itération à 76,91 % à la sixième itération). L'initialisation du dictionnaire permet une croissance plus rapide du nombre de messages traités automatiquement (de 58,74 % à la première itération à 89,76 % à la sixième itération).

On note que la majorité des ajouts au dictionnaire concerne la liste noire. Au cours des 6 itérations, une seule mention a été ajoutée à la liste grise.

5. Conclusion et perspectives

Nous avons proposé un protocole de construction d'un dictionnaire de médicaments à partir d'une base de données de médicaments. Notre approche a permis d'établir un classement des principes actifs les plus mentionnés sur les forums. Ce classement a servi de base pour un travail commun avec les centres de pharmacovigilance (CRPV) de Paris HEGP et de St-Étienne. Au cours de ce travail commun, les CRPV ont fourni des clés d'analyse (identification de vocabulaire patient, regroupement de médicaments, conjonctures événementielles) pour valider les résultats obtenus par exploitation de notre dictionnaire. L'application de cette méthode sur une base de données libre de droits, de qualité comparable à celle utilisée pour les expérimentations, ouvrirait le champ à la mise à disposition du dictionnaire ainsi constitué. Entre autres, le dictionnaire construit grâce à notre approche pourrait être exploité afin de de constituer des ensembles d'exemples positifs et négatifs pour paramétrer les outils d'apprentissage automatique afin d'améliorer la détection de syntagmes correspondant à des traitements médicamenteux.

Nos expérimentations ont porté sur une période et un nombre d'itérations réduits. Nous envisageons d'étendre la période étudiée et de faire varier la fréquence des itérations afin, entre autres, de vérifier les conséquences d'épiphénomènes tels que l'apparition ou la disparition de forums ou les épisodes médiatiques liés aux médicaments.

De même, nous avons choisi arbitrairement de sélectionner les 100 mentions de médicaments sans correspondance les plus fréquentes. Il serait intéressant d'étudier s'il existe un nombre optimal. Ce nombre nous a permis de remarquer que le nombre de messages impliquant les mentions de médicaments retenues diminue rapidement en quelques itérations. Il se pose alors la question d'une représentativité minimale pour les mentions que l'expert devra étudier.

Pour notre approche, nous avons utilisé une correspondance exacte (modulo une normalisation simple) entre les termes du dictionnaire et les termes issus des messages. Ce choix nous a permis d'identifier rapidement des variations orthographiques fréquentes qui sont ajoutées à la liste blanche. Cependant, on envisage de tester des mises en correspondances partielles, basées sur des mesures de similarité. Par ailleurs, le traitement actuel des caractères de Trademark est minimaliste et vise uniquement à permettre la mise en correspondance avec les entrées de notre dictionnaire. Cependant, la présence de ces caractères peut être exploitée pour cibler un ensemble réduit

de mentions de médicaments pour lesquels la probabilité d'identification est plus grande et pour lesquels la recherche peut être poussée davantage.

De même, l'exploitation d'une liste de laboratoires peut permettre l'identification de morceaux de syntagmes. En effet, il arrive que les auteurs des messages incluent le nom du laboratoire fabriquant le médicament évoqué. Lorsqu'il s'agit d'un princeps, le laboratoire n'est pas précisé dans le nom du médicament et les dictionnaires utilisés suivent cette convention. Par ailleurs, dans le cas des génériques, les médicaments fabriqués par certains laboratoires peuvent ne pas être présents dans le dictionnaire. Dans les deux cas, lorsque le nom d'un laboratoire est identifié, il peut être supprimé du syntagme et permettre l'identification du princeps ou du principe actif. Pour permettre l'identification des laboratoires, il est nécessaire de construire une base de données mettant en relation les laboratoires, la liste des dénominations sous lesquelles ils apparaissent et les médicaments produits par chacun d'eux. Nous nous sommes appuyés sur la Base de données publique des médicaments (BDM) pour commencer ce travail. La BDM est mise en œuvre par l'Agence nationale de sécurité du médicament et des produits de santé (ANSM), en liaison avec la Haute Autorité de santé (HAS) et l'Union nationale des caisses d'assurance maladie (UNCAM), sous l'égide du ministère des Affaires sociales et de la santé. Elle permet au grand public et aux professionnels de santé d'accéder à des données et documents de référence sur les médicaments commercialisés ou ayant été commercialisés durant les trois dernières années en France.

Au cours de nos expérimentations, nous avons remarqué que les noms des médicaments peuvent également comporter des informations de pharmacocinétique (ex : Varnoline continu). Ces informations ne sont pas toujours précisées par les usagers. Dans ce cas, les correspondances entre mentions de médicaments et dictionnaire ne sont pas trouvées. Le traitement des termes de pharmacocinétique permettrait d'améliorer les résultats de notre approche

Acknowledgements

Nous remercions l'équipe ISCOD de l'École des Mines de Saint Étienne pour l'extraction du contenu des forums, et le CNRS-LIMSI pour la détection des entités nommées dans les messages au moyen de méthodes d'apprentissage supervisé. Ce travail est supporté par les projets Vigi4MED et PHAReS de l'ANSM, d'une part, et Practikpharma de l'ANR, d'autre part.

References

Abou Taam, M., C. Rossard, L. Cantaloube, N. Bouscaren, G. Roche, et L. Pochard. « Analysis of patients' narratives posted on social media websites on benfluorex's (Mediator®) withdrawal in France. » *J. Clin Pharm Ther* 39, n° 1, 2014: 53-5.

Audeh, B., M. Beigbeder, A. Zimmermann, P. Jaillon, et C. Bousquet. « Vigi4Med Scraper : A Framework for Web Forum Structured Data Extraction and Semantic Representation. » *PLoS One*, n° 12(1), 2017.

Bagheri, H., I. Lacroix, E. Guitton, C. Damase-Michel, et J.-L. Montastruc. « Cyberpharmacovigilance : What is the usefulness of the social networks in pharmacovigilance ? » *Therapie*, n° 71(2), 2016: 235-9.

Eholié *et al*. « MuEVo, a breast cancer Consumer Health Vocabulary built out of web forums. » *International Semantic Web Applications and Tools for Life Sciences, SWAT4LS*. 2016.

Golder, S., G. Norman, et Y.K. Loke. « Systematic review on the prevalence, frequency and comparative value of adverse events data in social media. » *Br J. Clin Pharmacol*, n° 80(4), 2015: 878-88.

Hamon, Thierry, et Natalia Grabar. « Linguistic approach for identification of medication names and related information in clinical narratives. » *Journal of the American Medical Informatics Association*, 2010 : 549-554.

Husson, M.C. « Theriaque : independent-drug database for good use of drugs by health practitioners. » *Ann Pharm Fr*, n° 66(5-6), 2008: 268-77.

Lardon, J., R. Abdellaoui, F. Bellet, H. Asfari, J. Souvignet, et N. Texier. « Adverse drug reaction identification and extraction in social media : a scoping review. » *J. Med Internet Res*, n° 17(7), 2015.

Levin, Matthew, Marina Krol, Ankur M. Doshi, et David L. Reich. « Extraction and mapping of drug names from free text to a standardized nomenclature. » *AMIA Annual Symposium Proceedings*, 2007 : 438.

Morlane-Hondère, F., C. Grouin, et P. Zweigenbaum. « Représentation des informations textuelles pour la détection d'états pathologiques par apprentissage statistique. » *Journées Francophones d'Informatique Médicale*. Genève, 2016.

Nelson, Stuart J., Kelly Zeng, John Kilbourne, Tamy Powell, et Robin Moore. « Normalized names for clinical drugs : RxNorm at 6 years. » *Journal of the American Medical Informatics Association*, 2011 : 441-448.

Pahor, M., E.A. Chriscilles, J.M. Guralnik, S.L. Brown, R.B. Wallace, et Periugo Carbonin. « Drug data coding and analysis in epidemiologic studies. » *European journal of epidemiology*, 1994 : 405-411.

Palosse-Cantaloube, L., I. Lacroix, V. Rousseau, H. Bagheri, J.L. Montastruc, et C. Damase-Michel. « Analysis of chats on French internet forums about drugs and pregnancy. » *Pharmacoepidemiol Drug Saf*, n° 23(12), 2014: 1330-3.

Pereira, suzanne, Catherine Letord, Stefan Darmoni, et Elisabeth Serrpt. « Extraction des noms de médicaments dans les comptes rendus hospitaliers. » *Systè}mes d'information pour l'amélioration de la qualité en santé*, 2012 : 145-153.

Pimpalkute, Pranoti, Apurv Patki, Azadeh Nikfarjam, et Graciela Gonzalez. « Phonetic spelling filter for keyword selection in drug mention mining from social media. » *AMIA Summits on Translational Science Proceedings*, 2014 : 90.

Senger, Christian, Jens Kaltschmidt, Simon P.W. Schmitt, Markus G. Pruszydlo, et Walter E. Haefeli. « Misspellings in drug information system queries : characteristics of drug name spelling errors and strategies for their prevention. » *International journal of medical informatics*, 2010 : 832-839.

Sirohi, E., et P. Peissig. « Study of effect of drug lexicons on medication extraction from electronic medical records. » Dans *Biocomputing*, 308-318. World Scientific, 2005.

Tapi Nzali, Mike Donald, Sandra Bringay, Christian Lavergne, Thomas Opitz, Jerome Aze, et Caroline Mollevi. « Construction d'un vocabulaire patient/médecin dédié au cancer du sein à partir des médias sociaux. » *Ing{\'e}nierie des Connaissances*. 2015.

Zhou, Li, Joseph M. Plasek, Lisa M. Mahoney, Frank Y. Chang, Dana DiMaggio, et Roberto A. Rocha. « Mapping partners master drug dictionary to RxNorm using an NLP-based approach. » *Journal of biomedical informatics*, 2012 : 626-633.

Abstract

Drug identification within a textual corpus has several applications in various domains of pharmacology. Nevertheless, no dictionary of drugs is available for this identification in French. In this article, we propose a method to iteratively construct such a dictionary which is necessary to match patient vocabulary and expert vocabulary used in pharmacology.

Structuration de données pour un dictionnaire collaboratif hybride

Marie Steffens*, Kaja Dolar**, Noé Gasparini***

*Université d'Utrecht, Trans 10, 3512 JK Utrecht
m.g.steffens@uu.nl
**CREE, Inalco, 65 rue des Grands Moulins, 75013 Paris
dolar.kaja@gmail.com
***Institut international pour la Francophonie (2IF), 1C avenue des Frères Lumière, CS 78242, 69372 Lyon Cedex 08
noe.gasparini@univ-lyon3.fr

Résumé. Le *Dictionnaire des francophones* est un projet institutionnel de dictionnaire numérique qui agrège plusieurs ressources patrimoniales au sein d'une base de données relationnelle. Son modèle de données est guidé par l'ambition de proposer un outil collaboratif ouvert, permettant l'enrichissement par le grand public, tout en fournissant des données facilement réutilisables. Ce modèle permet de caractériser les formes et les définitions selon leurs lieux d'usages. Il intègre les traditions lexicographiques en les synthétisant et documentant afin de faciliter la contribution. Outre l'inscription dans la dynamique actuelle du web des données, le *Dictionnaire des francophones* propose également des forums de discussion qui intègrent des considérations sur la langue, sur les variations formelles ou l'étymologie des lexèmes et des notes sur l'usage. Des actions rapides inspirées des dictionnaires collaboratifs existants complètent le dispositif et facilitent la relecture *a posteriori* par les pairs.

1. Introduction

Se développant depuis une vingtaine d'années seulement, la lexicographie collaborative en ligne apparaît maintenant comme essentielle pour la lexicographie en général. Elle forme un champ de description linguistique vaste et diversifié quant aux modalités techniques du recueil de données, à l'éventail des types de données rassemblées et à leur représentation. Dans le cadre de

cette publication qui présente une nouvelle ressource collaborative en ligne, le *Dictionnaire des francophones* (DDF), nous adoptons la définition suivante : activité qui intègre les contributions d'une communauté et crée un espace virtuel dans lequel les contributeurs collaborent à la rédaction d'articles dictionnairiques (Dolar 2017a, Cotter & Damaso 2007, Meyer & Gurevych 2012, Granger 2012).

Le DDF est un projet de dictionnaire collaboratif numérique institutionnel. Ce projet est piloté par la Délégation générale à la langue française et aux langues de France (DGLFLF) et conçu par l'Institut international pour la Francophonie (2IF), composante de l'Université Jean Moulin Lyon 3. L'équipe de conception est composée des auteurs et autrices de cet article, soutenus par un conseil scientifique international.

Le DDF est un projet hybride qui présente de nombreux aspects novateurs par rapport aux ressources existantes, dictionnaires traditionnels et collaboratifs. Dans le présent article nous en exposerons trois principaux : le fonctionnement général du DDF qui intègre les ressources patrimoniales et une interface pour l'enrichissement collaboratif ; l'ontologie qui a été construite pour les fins du projet ; les divers modes de contribution qui se trouvent combinés dans le DDF.

2. Enrichir des ressources patrimoniales

2.1. Les ressources patrimoniales

Le *Dictionnaire des francophones* fait le choix de réunir plusieurs ressources patrimoniales figées en permettant leur enrichissement. Une démarche similaire a permis la constitution d'un substrat pour le *Wiktionnaire* francophone, grâce à l'intégration de la 8e édition du *Dictionnaire* de l'Académie française de 1932-1935 et du *Littré* de 1883. Ces ressources sont une richesse mais elles amènent des imperfections dues à leur ancienneté et à leurs méthodes de conception (Sajous *et al*. 2019). Les ressources intégrées dans le DDF sont également de diverses qualités.

2.2. L'Inventaire

La première ressource patrimoniale intégrée a été l'*Inventaire des particularités lexicales du français en Afrique noire* (IFA 2004). Il s'agit d'une compilation de onze études différentielles, donnant les correspondants en français

métropolitain des particularités locales. Chaque définition est précédée des zones où elle a été relevée, et les exemples proviennent des pays suivants : Bénin, Burkina Faso, Cameroun, Centrafrique, Côte d'Ivoire, Mali, Niger, République démocratique du Congo, Rwanda, Sénégal, Tchad, Togo.

Cet *Inventaire* est le fruit du travail de plus de vingt linguistes, de 1977 à sa première publication en 1983. L'Agence Universitaire de la Francophonie (AUF) a accepté de le placer sous licence libre afin de lui offrir une nouvelle diffusion, qui fut assez restreinte au moment de sa publication faute de soutien politique dans les pays concernés. Le contenu est richement décrit, mais parfois daté. Une actualisation est nécessaire.

2.3. Le Wiktionnaire

Le *Wiktionnaire* se définit lui-même comme un « projet lexicographique collaboratif accessible par internet, hébergé par la Wikimedia Foundation, sous licence libre, visant à décrire dans toutes les langues tous les mots »[1]. Seule la partie décrivant le français en français est intégrée dans le DDF. Cette ressource, rédigée par des lexicographes profanes (Vincent 2017) de formations inégales (Hanks 2012) a beaucoup évolué depuis 2004, en développant une approche descriptive à plusieurs mains, s'enrichissant de néologismes et de très nombreuses citations provenant de sources variées, mais elle demeure faible sur la description des usages. Ce n'est pas une ressource patrimoniale mais c'est une base libre de droit d'une qualité suffisante pour soutenir ce projet. Elle est constamment enrichie. Des copies de sauvegarde mensuelles devront donc être réinjectées régulièrement dans le DDF.

2.4. D'autres ressources

Le conseil scientifique du projet, présidé par Bernard Cerquiglini, a pour mission de réunir un maximum de ressources lexicographiques décrivant la langue française. D'autres ressources de cette teneur rejoindront la base de données du DDF par la suite, dont le Dictionnaire des belgicismes puis l'ensemble de la Base de données lexicographiques panfrancophone. Ces ressources seront clairement référencées dans le DDF et diffusées librement.

1 Contributeurs et contributrices du Wiktionnaire. Entrée « Wiktionnaire ». *Wiktionnaire, le dictionnaire libre* ; version du 12 sept. 2019 à 20 :58. https ://fr.wiktionary.org/w/index.php ?title=Wiktionnaire&oldid=27047574.

2.5. Licence libre

L'utilisation d'une licence ouverte pour les données s'inscrit dans la politique volontariste d'ouverture des données culturelles du Ministère de la Culture et de la Communication français[2] ainsi que dans la dynamique du plan S coordonné à l'échelle européenne pour l'accessibilité des données de la recherche publique[3].

2.6. Une base de données plutôt qu'un portail de ressources

Il ne s'agit pas de juxtaposer ces différentes ressources en construisant un index commun, comme le font le portail du CNRTL ou le site de l'Académie française mis en ligne en 2019. Il s'agit de mettre en réseau des dictionnaires au sein de bases de données décrites par un même modèle permettant une exploitation nouvelle, par un enrichissement interne collaboratif mais aussi par la réutilisation par d'autres en accédant directement aux données.

3. Une ontologie adaptée

Le DDF est d'abord une base de données structurée. L'étape la plus importante du projet pour permettre le balisage des ressources patrimoniales et la collecte de nouvelles informations lexicales a été la définition du squelette de la base de données et, dans un deuxième temps, des processus de contributions destinés à ajouter de la chair à ce squelette. Nous présentons ici le modèle de données.

3.1. Web sémantique

Les bases de données et les ontologies du Web sémantique (Antoniou & van Harmelen 2012) permettent de stocker les données sous la forme de graphes RDF (*Resource Description Framework,* W3C 2004a, Měchura 2016) organisés en triplets, rassemblant deux objets et leur relation, qui est réifiée et peut servir d'appui pour une requête au même titre que les objets. Il

2 https ://www.culture.gouv.fr/Presse/Archives-Presse/Archives-Communiques-de-presse-2012-2018/Annee-2014/Le-ministere-de-la-Culture-et-de-la-Communication-inscrit-son-action-dans-une-politique-volontariste-d-ouverture-des-donnees-publiques-culturelles

3 https ://www.coalition-s.org/

existe de puissants standards d'extraction, comme SPARQL (W3C 2013) qui permettent des requêtes complexes.

La structure sous forme de graphe de connaissance offre une grande capacité d'évolution et d'adaptation. Le langage de représentation des connaissances OWL (*Web Ontology Language*, W3C 2004b et 2012) permet des opérations logiques sur les données en RDF, à l'aide de classes, de propriétés et d'instances.

Ce balisage ne menace pas l'intégrité des ressources intégrées, les informations reprises dans les différents champs balisés, et dont la provenance est toujours indiquée, n'étant pas modifiables par les utilisateurs.

L'architecture de la base doit permettre de rationaliser les connaissances (méta-)linguistiques des locuteurs et de les intégrer à une structure informatique. L'objectif est de maximiser l'atomisation sémantique afin de favoriser l'enrichissement continu de la base de données et les renvois entre données, la structure en graphe permettant une progression non linéaire à partir de plusieurs points d'entrée.

3.2. Les modèles de données lexicographiques existants

Comme pour de nombreux autres dictionnaires numériques (ex. *Algemeen Nederlands Woordenboek*, Tiberius & Declerck 2017), c'est le modèle Ontolex Lemon, et son module lexicographique Lexicog, qui a été choisi comme base pour la structuration des données (W3C 2019). Il s'agit d'un cadre de modélisation de dictionnaires lisibles par machine, lié au Web sémantique (McCrae *et al*. 2017) et développé par un réseau international.

L'utilisation du modèle Ontolex Lemon permet de développer la base de données conformément aux nouvelles normes en matière de représentation des données linguistiques, intégrant des informations syntaxiques, morphologiques, sémantiques et sociolinguistiques dans les entrées lexicales.

Le souci de la communauté Ontolex pour la mise en réseau des données fournit également des outils pour la réutilisation des informations linguistiques existantes provenant d'autres ressources lexicales pour alimenter la base de données (Apel 2014), chaque nouvelle ressource patrimoniale intégrée dans le DDF devant être balisée d'après le modèle de données.

Le module Lexicog, dont la version définitive a été publiée en septembre 2019, apporte des nouvelles classes permettant de mieux décrire les structurations propres aux dictionnaires, tandis qu'Ontolex vise à la description de la langue.

Une ontologie connectée, Lexinfo, est également utilisée pour les inventaires de termes linguistiques qu'elle décrit.

3.3. Le modèle DDF

S'il est constitué majoritairement à partir d'Ontolex Lemon, le modèle de données du DDF (voir figure 1) en est une extension en raison des choix de modélisations propres au projet. D'abord, tout l'enjeu du DDF est de circonscrire l'usage des lexèmes du français dans une zone géographique donnée (ville, région, pays, continent). La propriété de localisation (« Place ») a donc été ajoutée tant à la description de la forme qu'à celle du sens, la variation diatopique touchant les deux : *soja* en Europe, mais *soya* au Québec, *savoir* utilisé dans le sens de « pouvoir » en Belgique, par exemple.

Ensuite, à chaque sens possible d'une forme peuvent être associées des marques sociolinguistiques d'emploi différentes, qui s'ajoutent au marquage diatopique. Le mot *sucre*, par exemple, pourra être étiqueté comme nom courant du saccharose dans les différentes variétés géographiques du français, comme nom courant de la sève d'érable au Québec ou comme terme technique vieillissant dans le domaine de l'électricité, synonyme de *domino* ou *serre-fils* en français de Belgique. Le modèle de données rend compte de ces possibles différences en optant pour un principe classique en modélisation de données : à chaque entrée dans la base de données correspond une forme – qui peut être une forme flexionnelle –, un sens et un ensemble de marques liées à la forme avec cette signification, les différentes entrées pouvant être reliées entre elles par la propriété "SenseRelation". L'inventaire des marques associées aux données et disponibles lors de la contribution vise à rendre compte de différentes traditions lexicographiques tout en étant structuré de manière à être accessible aux contributeurs. Les principales marques lexicographiques sont ainsi reprises, associées à des vocabulaires contrôlés, dans lesquels des choix ont été opérés (par exemple, le remplacement de la marque *populaire*, péjorative et désuète, par *très familier*).

Enfin, outre les valeurs contrôlées et les définitions et exemples rédigés, le modèle intègre des espaces de discussion spécialisés s'appuyant sur des ontologies existantes (autour de « sioc :Item ») et enrichie par des classes nouvelles pour les systèmes de votes, détaillés dans la partie suivante.

Chacun des champs du modèle de données est ouvert à la contribution des utilisateurs qui peuvent ajouter des données au contenu existant, sans le modifier. Nous allons maintenant présenter les différentes manières de contribuer au DDF.

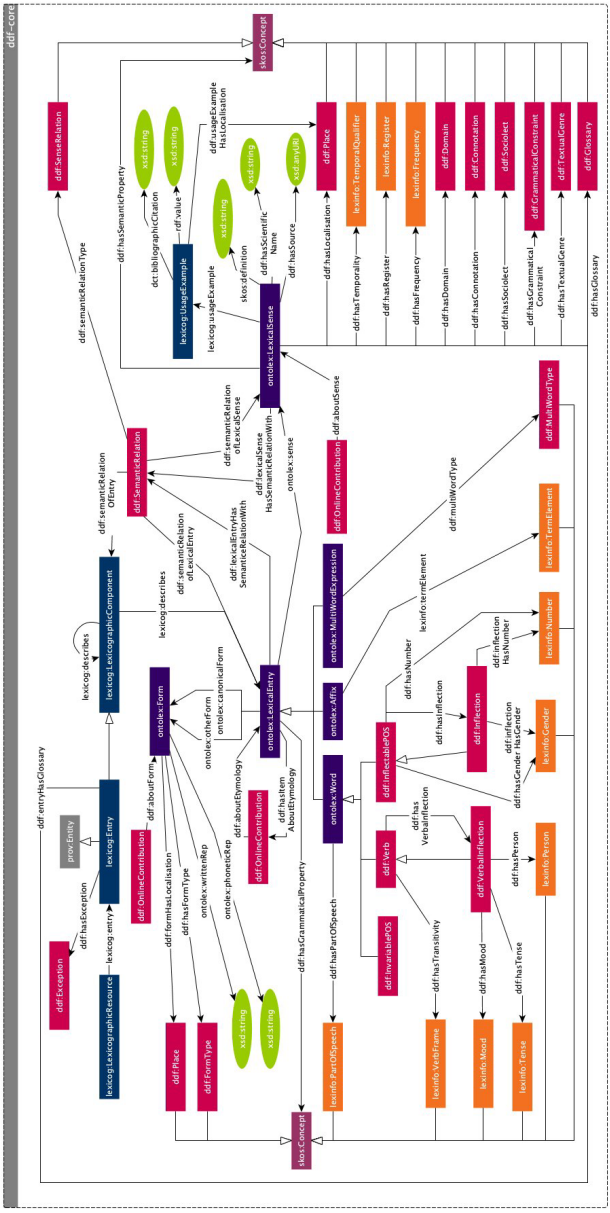


Figure 1 – Modèle de données du DDF

4. 1001 manières de contribuer

Les ressources collaboratives existantes se basent pour la plupart sur un seul mode de recueil de données alors que l'objectif du DDF est d'offrir la possibilité de contribuer de différentes façons et d'obtenir ainsi un maximum d'informations pertinentes. Ces informations peuvent être de natures différentes mais grâce à l'enrichissement fléché, les données soumises s'inscrivent dans le modèle de données. Les modes de contribution au DDF sont conçus pour faire la synthèse des bonnes pratiques des dictionnaires collaboratifs existants (pour un panorama, voir Dolar 2017b et Steffens 2017). La contribution au DDF est accompagnée aussi bien techniquement que pédagogiquement et elle est divisée en micro-tâches selon la nature de l'information apportée.

4.1. Enrichissement : formulaire et dialogue humain-machine

Le premier mode d'enrichissement est le formulaire de contribution (figure 2). Celui-ci peut donner lieu à :

- La création d'un nouvel article dictionnairique avec forme écrite, lieu d'usage, définition et exemple comme champs requis et d'autres informations non obligatoires.
- L'ajout d'une nouvelle définition ou d'un exemple à un article dictionnairique existant.
- De nouvelles informations à propos d'une entrée – ajout de différentes informations comme la catégorie grammaticale, les marques d'usage, les relations sémantiques et la prononciation à un article dictionnairique existant.

Ce type d'enrichissement est encadré et basé sur le modèle de données. Le but du formulaire de contribution est d'obtenir des informations lexicographiques structurées qui s'intègrent directement dans la structure dictionnairique. Il s'adresse à un public débutant, et introduit de courtes explications pour tous les termes techniques utilisés, tels que « étude et affinage du fromage » pour définir le domaine de la caséologie, par exemple.

Des éléments dynamiques vont enrichir ce formulaire, sous la forme de discussions avec la machine, dont les questions s'adapteront selon les réponses données. Ces outils permettront d'orienter un contributeur béotien sans introduire de vocabulaire technique et viseront dans un premier temps l'ajustement des relations sémantiques entre définitions, qui sont imprécises dans les ressources patrimoniales, et pourront être affinées au sein du DDF.

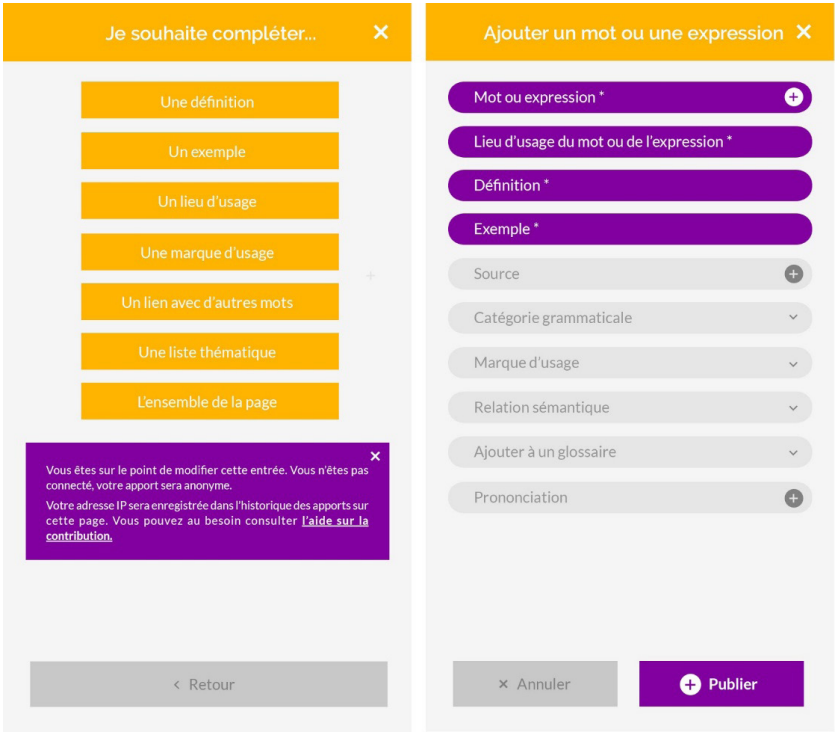


Figure 2 – Écrans de contribution (maquette préliminaire, version mobile)

4.2. Discussions de type forum

Le DDF accorde une importance particulière aux différentes discussions et négociations qui peuvent surgir autour d'une entrée (figure 3). Elles peuvent toucher divers aspects et sont organisées autour de trois grands thèmes :

- forme du mot, écrite comme orale transcrite,
- étymologie,
- usages, incluant les indications sur la politique linguistique, la norme et les marques dia-intégratives.

Les discussions prennent une forme libre, sans formulaire prédéfini. Elles permettent d'intégrer des informations qui ne s'insèrent pas dans le formulaire de contribution car ce ne sont pas des données structurées.

Figure 3 – Espaces de discussion (maquette préliminaire, version bureau)

4.3. Validation par les votes

Les contributions comme les discussions sont ouvertes aux votes. Le lectorat peut ainsi évaluer les contributions. Trois types de votes sont proposés : ✓ *Je valide* confirme la contribution et la fait remonter dans la liste des résultats de recherche ; [!] *Je signale* attire l'attention sur un problème, – *À supprimer* initie l'élimination de la contribution. Ces votes impliquent les usagers dans le processus de tri des apports.

Les informations contradictoires mais plausibles, telles que des étymologies concurrentes ou bien des formulations de définition qui décrivent le même sens mais rédigées différemment seront ainsi ordonnées sans que ne soient effacées les propositions ayant le moins de votes.

5. Pour terminer

Dans cette contribution, le DDF a été présenté sous l'angle général de son modèle de données et de ses principales fonctionnalités. D'autres angles auraient pu être abordés et le seront certainement dans le futur comme le positionnement de ce nouvel objet par rapport à la lexicographie différentielle, aux dictionnaires en ligne et aux ressources collaboratives. Nous pourrons également nous focaliser sur l'expérience utilisateur quant à la consultation et sur les mécanismes d'aide à la contribution. Ces dimensions sont essentielles dans la mesure où il apparaît clairement que dans les ressources lexicales collaboratives existantes, les choix techniques et didactiques ont un impact direct sur l'accessibilité, la qualité et la structuration des données.

Figure 4 – Un article du DDF (maquette préliminaire, version mobile)

Références bibliographiques

Antoniou, Grigoris & Frank van Harmelen. 2012. *A Semantic Web Primer*. Cambridge : The MIT Press.

Apel, Ulrich. 2014. "Linking a Dictionary to Other Open Data – Better Access to More Specific Information for the Users". In *Proceedings of the XVI EURALEX International Congress : The User in Focus*, edited by

Andrea Abel, Chiara Vettori & Natascia Ralli, 495-504. Bolzano : Institute for Specialised Communication and Multilingualism.

Contributors to the The OntoLex Lemon Lexicography Module Specification. 2019. *The OntoLex Lemon Lexicography Module*. Final Community Group Report 13 September 2019. Accessed Nov 13th 2019. https ://jogracia.github.io/ontolex-lexicog/

Cotter, Colleen & John Damaso. 2007. "Online Dictionaries as Emergent Archives of Contemporary Usage and Collaborative Codification". In *Queen Mary's OPAL #9 (Occasional Papers Advancing Language)*. London : University of London.

Dolar, Kaja. 2017a. *Les dictionnaires collaboratifs en tant qu'objets linguistiques, discursifs et sociaux*. PhD diss. Université Paris Nanterre.

Dolar, Kaja. 2017b. "Les dictionnaires collaboratifs non institutionnels dans l'espace francophone : éléments de typologie et bilan", In *Repères – Dorif* 14. Accessed Nov 13th 2019. http://www.dorif.it/ezine/ezine_articles.php ?art_id=380

Granger, Sylviane. 2012. "Introduction : Electronic Lexicography – from Challenge to Opportunity". In *Electronic Lexicography*, edited by Sylviane Granger & Magali Paquot, 1-14. Oxford : Oxford University Press.

Hanks, Patrick. 2012. "Corpus evidence and electronic lexicography", In *Electronic Lexicography*, edited by Sylviane Granger & Magali Paquot, 57-82. Oxford : Oxford University Press.

IFA = Équipe IFA. 1983. *Inventaire des particularités lexicales du français en Afrique noire*, AUPELF. Paris : Edicef.

Lexinfo. Accessed Nov 13th 2019. https ://www.lexinfo.net/

Meyer, Christian M. & Iryna Gurevych. 2012. "Wiktionary : A new rival for expert-built lexicons ? Exploring the possibilities of collaborative lexicography". In *Electronic Lexicography*, edited by Sylviane Granger & Magali Paquot, 259-291. Oxford : Oxford University Press.

McCrae, John., Julia Bosque-Gil, Jorge Garcia, Paul Buitelaar & Philipp Cimiano. 2017. "The OntoLex-Lemon Model : development and applications". In *Proceedings of eLex 2017*, edited by Iztok Kosem, Carole Tiberius, Miloš Jakubíček, Jelena Kallas, Simon Krek & Vít Baisa, 587-597. Accessed Nov 13th 2019. https://elex.link/elex2017/wp-content/uploads/2017/09/paper36.pdf

Měchura, Michal. 2016. "Data Structures in Lexicography : from Trees to Graphs". In *Proceedings of Recent Advances in Slavonic Natural Language Processing*, edited by Aleš Horák, Pavel Rychlý, Adam Rambousek,

97-104. Accessed Nov 13th 2019. http://www.lexiconista.com/raslan2016.pdf

Sajous, Franck, Nabil Hathout & Amélie Josselin-Leray. 2019. "Du vin et devin dans le Wiktionnaire : neutralité de point de vue ou neutralité et point de vue ?", *Éla. Études de linguistique appliquée*, 194, 147-164. https://www.cairn.info/revue-ela-2019-2-page-147.htm

Steffens, Marie. 2017. "Lexicographie collaborative, variation et norme : le projet 10-nous", In *Repères – Dorif* 14. Accessed Nov 13th 2019. http://www.dorif.it/ezine/ezine_articles.php?art_id=393

Tiberius, Carole & Thierry Declerck. 2017. "A *lemon* Model for the ANW Dictionary". In *Proceedings of eLex 2017,* edited by Iztok Kosem, Carole Tiberius, Miloš Jakubíček, Jelena Kallas, Simon Krek & Vít Baisa, 237-251. Accessed Nov 13th 2019. https://elex.link/elex2017/wpcontent/uploads/2017/09/paper14.pdf

W3C. 2004a. Accessed Nov 13th 2019. https://www.w3.org/TR/rdf-concepts/

W3C. 2004b. Accessed Nov 13th 2019. http://www.w3.org/2004/OWL

W3C. 2012. Accessed Nov 13th 2019. https://www.w3.org/TR/2012/REC-owl2-overview-20121211/

W3C. 2013. Accessed Nov 13th 2019. https://www.w3.org/TR/sparql11-overview/

W3C. 2019. "Contributors to the The OntoLex Lemon Lexicography Module Specification". In *The OntoLex Lemon Lexicography Module. Final Community Group Report*, 13 September 2019. Accessed Nov 13th 2019. https://jogracia.github.io/ontolex-lexicog/

Abstract

Dictionnaire des francophones is an institutional digital dictionary project that combines several traditional lexical resources within a relational database. The goal of the project is to create an open collaborative tool, allowing a wider public to contribute and enrich the content, while the data remains easily reusable. This model enables the possibility to relate the forms and the definitions to their respective geographical regions. In order to facilitate the contribution process, *Dictionnaire des francophones* integrates different lexicographic traditions by synthesizing and documenting them. *Dictionnaire des francophones* reflects the current dynamics in the field of web of data. It also offers discussion forums which might include remarks on language, formal variations, etymology of lexemes and usage notes. Rapid actions inspired by existing collaborative dictionaries have been developed; they facilitate the process of peer-review which takes place after the modifications are published online.

Articles courts

Creating a Terminological Resource: Importance and Limitation of Corpora

M. Ebrahimi Erdi*

*Autonomous University of Barcelona, Faculty of Translation and Interpretation, PhD Student
m.ebrahimi8938@yahoo.com

Abstract. The aim of this paper is communicating some preliminary results of the study of sensitive terms and concepts included in a corpus of political news extracted from two Persian pro-conservative and reformist news agencies. Frequency lists, concordances and n-grams will be analysed using SketchEngine and implications for translation will be drawn. As it is a corpus-based study, it is worth mentioning a couple of points about advantages and disadvantages of corpus.

1. Introduction

Corpus, defined as a collection of machine-readable texts, is a relatively new tool to help terminologists, among other linguistic experts, which is approaching to its truly significant status; although, it has its own shortcomings which can be overcome to a great extent. Given ever-growing advances in technology, specifically computational linguistics, collecting corpora of millions of words has become easier. Another advantage of using corpora in terminological researches is that you can make recent and up-to-date ones and tailor them according to specific criteria you have in mind or make sub-corpora of them based on different variables and check their significance or impact and/or the reliability based on statistics and quantitative information. Advances in corpus-based terminological tools are impressive and they are becoming more and more friendly-user.

Baroni (2009) highlights the importance of word frequency and considers it as an integral part 'in all branches of corpus linguistics' and asserts that it is 'what distinguishes corpus-based methodologies from other approaches to language' (p. 803). He also distinguishes between 'rank/frequency profile'

and 'frequency spectrum' which is related to the present paper. The former is specifically functional for investigating the characteristics of high frequency terms and the latter is specifically helpful to explore the 'properties of low frequency items' (p. 806).

2. Methodology

This paper is based on a corpus which was designed meticulously to be representative so that the results would be used for quantitative analyses and be generalizable to some extent, an issue highlighted by Biber & Jones (2009) among others. Amongst various categorizations of corpora, the one used in this paper is 'special purpose corpus' which is neither a general reference nor a monitor corpus, neither a special nor a subcorpus : "whenever the specific purpose for which the corpus is to be used … is the reason for creating or selecting the corpus", that corpus is referred to as 'special purpose corpus' (Pearson 1998, 48). The sampling method was 'stratified' in which, at first, the focused text categories (political texts) were selected and then, texts related to that category (political news of two news agencies broadly and clearly known as pro- conservative and reformist) were collected (*ibid*, p. 1288). As the focus is presidential election, they were filtered according to their relevance to this topic and also from one month before the election date, May 19th of 2017, during which campaigns and hot political debates were dominating the news agencies. The size of corpus is another important issue, which in case of this paper, it is almost two million words. It is worth mentioning that Persian is among less studied languages but is, according to Ostler (2008), 'probably upwardly mobile' (p. 458).

3. Results

Some main super-concepts previously selected on the basis of their relevance in the political context under investigation (e.g. power, peace, election, democracy) were studied using SketchEngine. Here are some of the results revealed in this study :

a) First of all, a point about corpora : Sometimes texts themselves or the purpose for which you want to exploit them are subject to copyright law ; if you are willing to have a corpus or a couple of subcorpora tailored according to specific criteria, it would become a time-consuming procedure especially if your corpus is of more than ten or twenty million words.
b) There were numerous lexical gaps due to different political system of

Iran and they would cause troubles during translating from Persian to English such as *estekbar* (countries which are bigoted and don't follow righteousness), *nezam-eh solteh* (domineering countries which don't respect the autonomy of the others), *basirat* (a special consciousness to recognize enemy, even the national one, from friend or wrong from right), *delvapasan* (those who are worried which is a sarcasm usually used by reformists referring to critics expressed by conservatives) and etc. .

c) There were very important political terms among words with low frequency (2 to 5 times of occurrence); for instance : *nofouz* (penetration), *fetneh* (sedition), *tote'eh gar* (revolt) and so on which proved that low frequency terms should not be ignored at all.

d) N-grams depicted to be functionally crucial because of their information-loaded aspects which can be ignored if just word list or concordances of higher frequency were investigated.

e) Last but not least important point was usefulness of corpus linguistics (CL) tools and techniques for the study of less studies languages, on the one hand, and bringing deep-structured and hidden semantic networks to surface.

Bibliography

Adel, Annelie and Randi Reppen Eds. 2008. *Corpora and discourse : The challenges of different settings*. Amsterdam/Philadelphia : John Benjamins.

Baroni, Marco. 2009. "Distributions in text". In *Corpus linguistics : An international handbook* edited by Anke Ludelingand Merja Kyto, Vol. 2, 803-822. New York, Berlin : Walter de Gruyter.

Baroni, Marco and Stephan Evert. 2009. "Statistical methods for corpus exploitation". In *Corpus linguistics : An international handbook* edited by Anke Ludelingand Merja Kyto, Vol. 2, 777-803. New York, Berlin : Walter de Gruyter.

Biber, Douglas. 1993. "Representativeness in corpus design". *Literary and linguistic computing* 32 : 243-257.

Biber, Douglas and James K. Jones. 2009. "Quantitative methods in corpus linguistics". In *Corpus linguistics : An international handbook* edited by Anke Ludelingand Merja Kyto, Vol. 2, 1286-1304. New York, Berlin : Walter de Gruyter.

Gulinski, Christian and Klaus-Dirk Schmitz. 1996. *Terminology and knowledge engineering : Proceedings of 4th international congress on terminology and knowledge engineering*. Germany : INDEKS Verlag.

Ludeling, A. & Kyto, M. (Eds.) (2008). *Corpus linguistics : An international handbook* (Vols. 1-2). New York, Berlin : Walter de Gruyter.

Ostler, Nicolas. 2008. “Corpora of less studied languages”. In *Corpus linguistics : An international handbook* edited by Anke Ludelingand Merja Kyto, Vol. 1, 457-483. New York, Berlin : Walter de Gruyter.

Pearson, Jennifer. 1998. *Terms in context*. Amsterdam/Philadelphia : John Benjamins.

Wynne, Martin. 2008. “Searching and concordancing”. In *Corpus linguistics : An international handbook* edited by Anke Ludelingand Merja Kyto, Vol. 1, 706-737. New York, Berlin : Walter de Gruyter.

Résumé

Le but de cet article est de communiquer quelques résultats préliminaires de l’étude de termes et concepts sensibles inclus dans un corpus de nouvelles politiques extrait de deux agences de presse persans pro-conservatrices et réformistes. Les listes de fréquences, les concordances et les n-grammes seront analysés à l’aide de SketchEngine et les implications pour la traduction seront tirées. S’agissant d’une étude basée sur un corpus, il convient de mentionner quelques points concernant les avantages et les inconvénients du corpus.

Company-speak : The glue of corporate culture

Benedikt Jankowski, MA

benijankowski@gmail.com
Haller Strasse 77, 6020-AT Innsbruck

Abstract. The present case-study deals with the importance of Company-Speak within corporate culture. Whilst Company-Speak often is negatively called jargon, many companies do not see the use of such vocational sociolects for its corporate culture as well as communication. Unifying the entire vocabulary of a company can be problematic for departments. Therefore, a terminological solution is required such as creating a framework in which Company-Speak can evolve but still can be controlled. This task is made for terminologists with their precise as well as thorough way of working. With such a framework and terminology, departments know when to use which term when communicating with another department. Furthermore, it promotes the corporate identity, as employees have the chance to form it. For the correct use, writing programmes which control the terminological and scriptoral use can be implemented to guarantee a good communication.

1. Introduction

Due to miscommunication as well as lack of motivation, companies around the globe lose millions of Euros. It is a hidden leak, which barely gets detected by the leading figures of a company. According to experts, German companies lose 1 billion Euro per year due to miscommunication (Dezes, 2009 :45).

What has this number to do with Company-Speak, one may ask. As Company-Speak (an vocational sociolect) is part of Corporate Language, it is very well connected with the topic. Company-Speak possesses several tasks within a company, among which are a more efficient communication as well as gluing together Corporate Identity. Hence, Company-Speak takes over the same task as socio- and dialects from our daily lives ; namely to communicate and to unite.

However, Company-Speak is often seen as *jargon*. Possessing such a bad reputation, many companies try to avoid developing occupational sociolects. Moreover, they tend to unify the company's entire vocabulary, which can be problematic as well.

Thus, the questions dealt with in this article are : What is Company-Speak ? What are its tasks ? And how can companies benefit from using Company-Speak ?

2. What is Company-Speak ?

As previously mentioned, Company-Speak can be seen as occupational sociolect. Therefore, the term sociolect must be defined first. A very accurate definition of the term comes from Löffler (1994):

> *"Wherever there is a group, characterised by social, vocational, technical, status and reputation based aspect, which also shows a linguistic or grammatical-lexical-intonational characteristics, these varieties should be called 'sociolectal' or 'sociolect'."*

Thus, as Company-Speak develops itself within a vocational environment and has its own linguistic characteristics, it can be declared as sociolect. Examples of such a term can be one word which describes an entire process or even product names.

A study over a period of 11 years conducted by Kotter and Heskett (1992) concluded that companies with performance-enhancing cultures show significantly better results in stock price growth (901 %) compared to the ones with no such efforts (74 %). This can be seen as a reaction of customers as well as investors ; but what is a corporate culture ? The notion of corporate culture is composed by the company's branch (customers), its products (quality vs. quantity), marketing (slogans), PR (perception) as well as internal and external communication (company-speak).

Even though all aspects of corporate culture are important, company-speak signifies the most important one, as it can influence all the other aspects of corporate culture. Company-speak has the crucial task of gluing all the important parts of a company and its background culture together. Employees should ideally work together as a unit and most importantly, be able to identify themselves with their company. The motto should be as the founder of Virgin Airlines, Richard Branson (Boyle, 2018), quoted :

"Clients do not come first. Employees come first. If you take care of your employees, they will take care of the clients."

3. Company-Speak is not Jargon

In a recent study, de Vecchi (2018) claims that company-speak is often pejoratively called *jargon*. It has been characterised as a not-precisely defined notion and as a positive mask used to rename negative concepts, for instance, problems are called "opportunities". But is this actually the purpose of company-speak?

On a closer look, de Vecchi claims that company-speak underwent several changes over the course of the 20th century. During and after the industrialisation, companies wanted to increase their numbers within a short amount of time and therefore, employees had to be seen as well-functioning machines. It is also the time, when the term *human resources* occurred first, which has generally not been seen as a positive term. However, there have been positive ideas, as the one of sociologist Elton Mayo (Spicer, 2017). In the 1930's, he claimed that the human aspects are more important than the environment at the working place. Unfortunately, this movement experienced a turn in 1984 (Spicer, 2017), when Charles Krone introduced the so-called *Kroning*, which defines the denounced *jargon* of today.

4. Company-Speak without terminology

As mentioned before, Company-Speak can be seen as the most important tool or aspect of corporate culture. However, Company-Speak often cannot be generalised on an entire company. The reason for that is that departments do not have that much direct contact with each other. This leads to variations within a company, which can cause severe communication issues.

Reason for that is pretty obvious. It is the same situation as with socio- and dialects in general. If we do not know a word it is very difficult to understand its meaning. Therefore, it can be very counterproductive or even dangerous for a company to risk such misunderstandings. As a solution, companies try to unify the vocabulary in order to avoid such mistakes.

However, terminologies, which could help solve such problems, are still not common within the business world. A study with 940 German companies, conducted by Schmitz and Straub (2010), showed that only 1 % is thoroughly investing into terminology. Moreover, 22 % claim that there are lots of com-

municational mistakes due to different names for products and 62 % consider terminology as important. This means that companies want to avoid communication errors but are still not motivated enough to invest into terminology. Why is this so ?

5. Why terminology is still barely seen amongst companies

Investing into terminology is combined with costs and time. Thus, companies do not like to invest either of them in order to be more efficient in terms of communication. Reason for that is not only the investment factor itself but also the shareholders.

They invested into the company and therefore want to see good numbers. Every year, they expect better results than in the year before and companies are forced to act. However, misunderstandings within the company itself is part of the reason, why it is hard to produce better numbers than before.

Another reason why terminology is not a common topic within the business world is the detection of miscommunication and problems itself. Measuring communication problems within a company is not a simple task, as they are not shown within accounting. Managers only see that departments produce certain numbers. Misunderstandings are not amongst them, which can be very dangerous in the long run. More and more, vast sums of money disappear without any obvious reason.

6. Combining Company-Speak with terminology

In order to make Company-Speak effective, terminology is the best tool to gain the best out of both issues ; communication problems as well as diversity. Company-Speak should not be forbidden by companies, as they are vital for corporate culture and corporate identity. However, if not standardised, Company-Speak can become the worst nightmare for corporate communication.

Therefore, terminologists should be responsible for the standardisation process. Possessing various expressions for the same product or process, the task of terminologists needs to be extended. They need to filter out sociolect expressions and standardise them as well. Hence, departments are aware of the other departments' vocabulary. This means, that there is not only one but several columns which present the different expressions of the departments.

6.1. Why not unifying the company's vocabulary?

The question arises, why a unification of all terms would not be the best solution. Two reasons come to mind when faced with this question. First, departments become too restricted when dealing with language. Creativity, which is an important aspect of Company-Speak will be lost entirely. Thus, motivation as well as effectiveness will decline drastically.

Second, some expressions are simply not good for use in certain departments, as they are more effective working with categorised names for products or processes. In research and development, product names often only consist of letters and numbers. Such names, however, are not good for marketing or other departments dealing with customers. Therefore, different expressions for a term are often the only solution.

7. Taking Company-Speak a step further

Company-Speak can even be taken to another level by motivating departments and employees to be creative. This includes renaming departments entirely as well as product names. Such a use of language can almost be compared to the one used in the military with names of secret operations or units.

Terminologists can help this process by creating a framework, in which departments are allowed to move with their creativity. They decide, what the boundaries are and how the framework relates to the aimed or existing corporate culture should look like. Exchanging traditional names of departments such as HR or RD can be an energising process for employees and departments.

7.1. Example with the start-up First Class Luxury Company

As my Master's thesis, I conducted a case study in which I created exactly such a framework within the start-up First Class Luxury Company. My task was to create a foundation of corporate culture with the help of terminology.

Step one was to analyse the desired goal of the company. The company backed its culture on the Habsburg, Ottoman as well as Byzantine Empire, which inspire the products as well as their crafting. Moreover, it is a luxury company, which asks for a strong background culture in order to be successful.

Thus, I used these cultures to create my terminological framework. I created a corpus to filter out all significant terms and prepared the official terminology of the company, meaning that I created the company's terminology for external use. The terms are the ones used for external communication with customers and co-operators.

Furthermore, a boost of morale was a positive side result. Employees were proud to be called Lions or other animals. Corporate culture developed basically itself, as employees were looking voluntarily for new cultural terms on a regular basis. Moreover, employees started to know the background cultures better and thus, became experts regarding the company's cultural history.

In order to guarantee the permanent and correct use of terminology, language programmes, such as TextLab, can be used. They can include a terminology and they even supervise the style of writing by checking its readability. When a term is used wrong, the programme tells you the correct use.

7.2. Programmes to support the use of Company-Speak

Within companies, not every employee is as trained to use terminology and to learn vocabulary as terminologists. Therefore, programmes can be used to simplify the process of communication. They can be fed with the terminology to support the author when using the wrong term and they help in terms of readability.

This happens with the help of so called readability indices, which exist for several different languages with different parameters. Depending on the audience, the index can be adapted meaning that 1-30 is very difficult to read and only for experts, 31-59 is difficult to read and 51-100 is easy to read.

One such programme is TextLab, which was developed by H&H Communication Lab GmbH (Haug & Haseloff, 2018). It is very successful with more than 30 000 users worldwide. Users can even select categories such as which department they are writing to or type of text such as Marketing or PR. Using such programmes can facilitate the writing process immensely.

Conclusion

Company-Speak is very important regarding corporate culture. It can be very useful when standardised to provide enough language-related support in communication. Therefore, terminologists are necessary in order to locate

terms of Company-Speak value and include them into the company's terminology.

Hence, departments are able to use the ideal form of products and processes within their department and know the correct counterpart of other departments, which is often the case.

Using Company-Speak as an advantage in terms of communication, creativity and corporate identity, terminologists are responsible for creating a cultural framework for the naming process within the company. Such a framework can boost morale and help employees to connect with their company.

However, learning terminology and vocabulary is not always an easy and beloved task among employees. Therefore, programmes, such as TextLab, can be included. They simplify the writing process by helping with the correct use of terminology as well as with the readability of the text.

Accepting and even supporting Company-Speak can be used as an advantage and by implementing programmes, the company can be sure that employees are using the correct term. As a result, companies can save vast amounts of time as well as money in the long run.

References

Boyle, Charlie. 12 Jan. 2018. *"Clients do not come first. Employees come first. If you take care of your employees, they will take care of the clients."*. https://www.linkedin.com/pulse/clients-do-come-first-employees-you-take-care-your-charlie-boyle.

Haug, Oliver/ Haseloff, Anikar. 2018. *"Corporate Language: Unternehmenssprache verständlich gestalten, effektiv steuern und praxisnah umsetzen." Stuttgart, Kohlhammer.* Kotter, J. P., Heskett J. L. (1992). *Corporate culture and performance*. New York: Free Press.

Löffler, H. (1994) *Germanistische Soziolinguistik*. 2nd revised edition. Berlin: Schmidt.

Spicer, A. (23 Nov. 2017), *From inboxing to thought showers: how business bullshit took over, Vacuous management-speak is easily laughed off – but is there a real cost to talking rubbish*?, https://www.theguardian.com/news/2017/nov/23/from-inboxing-to-thought-showers-how-business-*bullshit-took-over.*

Schmitz, Klaus-Dirk/ Straub Daniela (2010) *Erfolgreiches Terminologiemanagement im Unternehmen; Praxishilfe und Leitfaden:*

Grundlagen, Umsetzung, Kosten-Nutzen-Analyse, Systemübersicht. Stuttgart, TC and more GmbH.

Vecchi, D. (2018). *Company-speak, organisation-speak*. In Humbley, J., Budin, G. & Laurén, C. (eds.), Languages for Special Purposes. An International Handbook, 279-288, Berlin, Boston : de Gruyter.

Poésie (al-)chimique. Comment approcher le langage de l'alchimie néo-latine du XVII^{e} siècle à travers un thesaurus Semantic Web ?

Sarah Lang*

*Zentrum für Informationsmodellierung, Elisabethstraße 59/III, 8010 Graz, Austria
sarah.lang@uni-graz.at

Abstract. La méthode de communication alchimique est l'analogie. Il se pose la question comment on peut bien représenter le savoir alchimique dans le domaine numérique et trouver un moyen pour relier de manière fiable les mots codés des documents historiques à cette représentation de savoir. Ceci peut être obtenu en créant un thésaurus systématique classifiant les concepts qui sont représentés par les mots des texts historiques.

1. Théorie des *Decknamen*

La méthode de communication alchimique est l'analogie. Surtout dans la communication publique (imprimée), l'alchimiste ne peut pas parler ouvertement afin de ne pas rapporter les secrets alchimiques au public ne pas digne de les connaître (cf. Principe 2013). Cette présentation discute comment on peut bien représenter le savoir alchimique dans le domaine numérique et trouver un moyen pour relier de manière fiable les mots codés des documents historiques à cette représentation de savoir. Ceci peut être obtenu en créant un thésaurus systématique classifiant les concepts qui sont représentés par les mots des texts historiques.

Dans ce projet, il s'agit de chercher une manière de représenter le savoir alchimique historique qu'on trouve dans les sources historiques de l'alchimiste Michael Maier (1568-1618). Ici, il s'agit de rendre explicites les connaissances tacites qui se trouvent, implicites, dans les textes. Comme la relation entre rouge et or mentionné auparavant.

Aujourd'hui, l'analogie alchimique des fameux *Decknamen* (mots symboles) est souvent comprise comme moyen à coder un message dont le résultat ne peut plus être résolu de manière sûre. Cela était à tort, puisque la recherche contemporaine a eu du succès en recréant des recettes alchimiques dans des laboratoires chimiques (cf. Principe 2013). La logique RDFS, par exemple, permet de représenter des relations analogiques et alors, il se demande si les secrets alchimiques peuvent être résolus en se servant des méthodes des Humanités Numériques.

Pour donner un exemple du parler alchimique, la relation implicite omniprésente entre le rouge et l'or dans le langage alchimique (cf. Werthmann 2011) peut être rendu explicite ainsi par des triples Semantic Web :

PhilosophersStone hasColor red.

colorRed hasChemicalProperty tinctura (la teinte).

tinctura givesPhysicalProperty citrinitas (le jaune).

Gold hasColor citrinitas.

Ainsi, dans un texte où le sang du Christ (ayant la couleur rouge) peut être relié à la pierre philosophale, comme il est typique du symbolisme alchimique. Tenant d'autres signes en compte, nous savons donc qu'un texte alchimique pourrait parler de la pierre philosophale qui est créée symboliquement par le sacrifice du Christ. La relation représentée sert aussi à attirer notre attention sur le fait que partout où on parle de quelque chose qui est rouge, il se peut qu'en réalité, cela renvoie à l'or.

2. Application

Le fait de communiquer par analogies a pour résultat que le signifiant et le signifié sont reliés par des qualités ou relations en commun. Un thesaurus formalisant les qualités et relations de ces concepts aide à les résoudre. SKOS-XL permet de définir de manière très exacte les relations entre concepts et différents labels. La présentation a discuté des particularités dans la représentation du savoir dans un domaine historique qui mélange les sciences avec de la poésie et réception de mythologie antique. En particulier, comment représenter les différents contexts de quelques concepts (mythologique, théories des humeurs, colours, etc.). Après avoir réalisé une première représentation des concepts alchimiques dans un thésaurus, il s'agira d'essayer de désambiguïser les sens des mots dans les contextes historiques.

References

Assmann, Aleida. 2015. *Im Dickicht der Zeichen.* Berlin.
Collins, Harry. 2011. *Tacit and Explicit Knowledge.* Chicago.
Maier, Michael. 1614. *Arcana Arcanissima.* London.
Maier, Michael. 1618. *Viatorium.* Oppenheim.
Obrist, Barbara. 1993. "Les rapports d'analogie entre philosophie et alchimie médiévale". In *Alchimie et philosophie à la Renaissance,* edited by Jean-Claude Margolin and Sylvain Matton, 43-46. Paris.
Polanyi, Michael. 1966. The Tacit Dimension. London/Chicago.
Principe, Lawrence M. 2013. *The Secrets of Alchemy.* Chicago.
Werthmann, Rainer. 2011. "Das Rot aus dem Gold - das Gold aus dem Rot. Glauber und das Goldrubinglas", in Stephanie Nomayo (éd.): *Johann Rudolph Glauber. Alchemistische Denkweise, neue Forschungsergebnisse und Spuren in Kitzingen*. Kitzingen am Main, 259-276.

Abstract

The method of alchemical communication is analogy. The question is how alchemical knowledge can be represented digitally and find a reliable means of linking encoded words from historical documents to this digital knowledge representation. This can be achieved using using a thesaurus which systematically classifies the concepts represented by the words in the neo-latin texts.