Normalizing speech transcriptions for Natural Language Processing

Anne Dister — Université de Louvain et Facultés universitaires Saint-Louis — *anne.dister@uclouvain.be*

Matthieu Constant — Université Paris-Est — *mconstant@univ-mlv.fr*

Gérald Purnelle — Université de Liège — *gerald.purnelle@ulg.ac.be*

Morphosyntactic tagging and syntactic parsing are key parts of Natural Language processing. Many systems now reach exploitable results for written French texts (Véronis, 2000; Clément, 2001), but there were rare attempts to automatically annotate spoken textual data (see though Mertens, 2002; Valli et Véronis, 1999). Indeed, existing software are inadequate to analyse texts transcribed from speech and face specific problems, all related to the nature of the data:

- for theoretical reasons (Blanche-Benveniste and Jeanjean, 1987), transcriptions of speech do not contain punctuation marks; nevertheless, most of the tools in Natural Language Processing are based on these marks in order to perform an initial segmentation of the text;
- texts include meta information that does not need linguistic analysis (e.g. names of speakers, information on enunciation context)
- texts contain lexical particularities specific to speech
- finally, spoken texts are full of disfluencies, i.e. locations in the speech flow where the syntactic linearity is broken because it is interrupted for some time at a particular position on the syntagmatic axis: e.g. overlapping statements, word fragments, self-correction...

Although spoken corpus annotation does not seem to be a specific problem (Benzitoun *et al*., 2004) given the fact that there is no

grammar for spoken to be opposed to a grammar for written (Blanche-Benveniste *et al*., 1990), the problems listed above need to be solved to obtain effective annotation systems. Indeed, we will see that speech transcriptions form a "new type" of texts with specificities that have to be taken into account by the analysers. In particular, disfluencies constitute a practical issue for automatic analysis of spoken texts, as many authors have already noted by reference to different languages (Adda-Decker *et al*., 2003; Bénard, 2005; Benzitoun, 2004; Benzitoun *et al*., 2004; Garside, 1995; Guénot, 2005; Nivre and Grönqvist, 2001; Oostdijk, 2003; Valli et Véronis, 1999, etc.). The specificities of spoken language considerably reduce the performance of tools initially implemented for standard written texts. The solutions adopted by the researchers in order to deal with the disfluencies are strongly dependent on the chosen approach, the task to be carried out and the tools that are being used.

Our solution consists in implementing a preprocessing module which normalises spoken texts in order to make them compatible with standard NLP tools. On the basis of a corpus of almost 500.000 words from the textual data bank of spontaneous spoken French of VALIBEL[1] research centre, we have especially studied four types of disfluencies: repetition, word fragments, immediate self-correction and the word *euh*, called "filled pause". We have shown the regularity of these phenomena in the corpus (which are the words, part-of-speech and syntactic structures involved), and the numerous interactions between them. In this paper, we will show how these four types of disfluencies were automatically identified in texts. The principle we used was to annotate the part of the disfluency called *reparandum* (according to the terminology in Shriberg, 1994), in order to keep only the *repair* part (see below).

The paper is organised as follows. Firstly, we describe the specificities of the spoken corpus used. Then, we formalise speech particularities in order to be easily identified by the preprocessing tool. Finally, we present the resulting tool and its outputs.

[1] VALIBEL for VAriétés LInguistiques du français en BELgique: http://www.uclouvain.be/valibel.html

1. Speech transcription

Speech transcription is not an easy task. Blanche-Benveniste and Jeanjean (1987) showed with French examples that it does not only consist for the transcribers in putting into written form what he/she hears. Transcribing requires making choices at different levels (what to be transcribed? How to transcribe?). It involves an analytic and interpretative work that has been called "translation" (Cheepen, 1995), "heuristic representation" (Mondada, 2000) or "deformation" (Bally, 1935).

From its creation in 1989, the VALIBEL research centre, which constructs and exploits large spoken corpora, established explicit transcription guidelines (Dister *et al.*, 2006). They follow three main principles: use of standard spelling, no use of punctuation marks and emphasis on the speech specificities in the data (Dister and Simon, 2007).

1.1. Use of standard spelling

The transcriptions we deal with rigorously follow standard spelling conventions. Therefore, there are no graphical deformations that would consist in making a strict correspondence with the pronunciation[2], as it can be frequently found in noble dialogues. From a spelling perspective, transcriptions cannot be distinguished from standard written French: no wild elisions (e.g. *j'suis* standing for *je suis* (*I am*), *p'tit* for *petit* (*little*)), no graphical "monsters" (e.g. *ché pas* pour *je sais pas* (*I don't know*), *pasque* for *parce que* (*because*)). All lexemes used can be found in standard texts as listed in language references like dictionaries. From a Natural Language Processing perspective, words are analysed on the basis of lexical resources containing them.

1.2. No punctuation marks

Usually, corpora of speech transcriptions built for linguistic research do not contain punctuation marks. Indeed, there exist no strict correspondences between prosodic phenomena and written punctuation. A short pause does not always correspond to a comma in

[2] When necessary, transcribers can add pronunciation information in meta-tags.

written texts. Furthermore, a longer pause does not systematically imply the use of stronger punctuation marks.

Blanche-Benveniste and Jeanjean (1987: 139) plead for speech transcriptions with no punctuation marks. They argue that their use implies that transcribers suggest an analysis before having performed it. This is why the notion of sentence has been abandoned in the studies on speech production. Therefore, we might wonder what minimal unit is required by NLP tools that need sentence segmentation.

Although punctuation marks are not used to annotate the corpus, silent pauses are inserted to help reading. There are three degrees of pause that were subjectively assigned depending their duration: / (short pause), // (long pause), /// (silence).

1.3. Emphasis on spoken specificities

Disfluencies

Studies on spoken language highlighted specific phenomena that are generally called *disfluencies*. They correspond to locations of the speech flow where the linearity is broken because it stops for some time on the syntagmatic axis. We name this way punctuation words (*ben, bon...*), the filled pause *euh* (*uh*), repetition of words or word sequences (cf. 2.1), immediate self-corrections such as *le la fille* (*the the girl*, cf. 2.2), word fragments (transcribed with the slash symbol /: *à Bru/ à Bruxelles* (*in Bru/ in Brussels*), cf. 2.3), etc.

Disfluency transcription demands a careful attention from transcribers, in order to write down such phenomena that are usually filtered by an ordinary listening. Indeed, they are so frequent in spontaneous speech that we unconsciously tend to ignore these marks.

Speaking slots

The sound continuum, that has become linear with the transcription, is divided into speaking slots, defined by the change of speaker. In our transcriptions, the sequencing of speaking slots is presented horizontally: words of the speakers succeed to each other top-down on the screen. Each paragraph represents the intervention of a speaker.

This organisation that Edwards (1995) calls vertical format is coherent with our reading habits: we start reading from the top of the

screen and what we read before occurs in time before we read after. Theatre texts adopted this format from a long time.

Overlapping statements

In standard spontaneous conversations, it is very frequent that two or more persons speak at the same time. We therefore have overlapping speech statements. In our transcription convention, symbols | and - delimit overlapping segments (|- for the beginning of the overlapping and -| for the end). For instance, in the following transcription

L1 je le connais |- depuis longtemps
L2 oui tu -| l'avais rencontré à mon mariage

L1 I know him |- for a long time
L2 yes you -| had met him at my wedding

speaker L2 starts to speak while a speaker L1 is already speaking; L2 continues and L1 stops.

However, it can happen than the second speaker starts to speak during the first speaker' speech, but the latter keeps on speaking after the overlap. We then have an internal overlapping segment that is transcribed as following:

L1 je l'aime |- vraiment beaucoup <L2> je sais -| ce chercheur

1.4. Transcription example

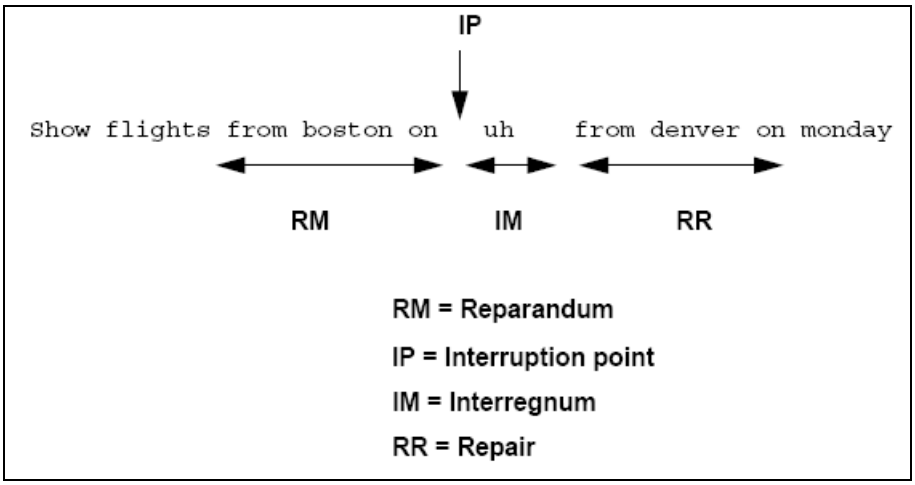
ileGF0 une une trémie / ça veut dire quoi
ilePA2 une trémie justement une trémie i/ |- c'est une < ileGF0> oui -| un tunnel une trémie chez nous c'est / c'est le c'est c'est ce qu'on appelle un tunnel
ileGF0 ah d'accord
ilePA2 hein |- mais < ileGF0> mm -| une pet/ un petit tunnel qui n'est pas très long
ileGF0 mm
ilePA2 or une trémie euh grammaticalement c'est une chose qui s'en/ qui s'enfonce plutôt dans la terre

ileGF0 a a hopper / what does it mean
ilePA2 a hopper precisely a hopper i/ |- it's a < ileGF0> yes -| a tunnel a hopper in our country it's / it's the it's it's what is called a tunnel
ileGF0 oh well !
ilePA2 eh |- but < ileGF0> mm -| a lit/ a little tunnel which is not very long

ileGF0 mm
ilePA2 now a hopper er grammatically it's a thing which pen/ penetrates in the earth

2. Identifying disfluencies

Shriberg (1994: 7-9), following Levelt (1989), represented the disfluent sequence by splitting it into four distinct elements corresponding to three regions:



• ***reparandum***: the *reparandum* (*RM*) is the part produced by the speaker that will not be kept and that will be replaced later by the *repair*;

• ***interrupting point***: the interrupting point (IP) is the moment of the speech that coincides with the end of the *reparandum*. It has an empty textual content;

• ***interregnum***[3]: the *interregnum* (*IM*) is the region that begins at the end of the reparandum and ends at the begining of the repair. It can contain an *editing term*, i.e. a silent pause, a filled pause, or several attempts of unachieved reformulation;

• ***repair***: the *repair* (*RR*) indicates the correction of the *reparandum*.

[3] The *interregnum* corresponds to Levelt's editing phrase.

For a long time, researchers showed regularities constraining disfluencies, especially in English (Blankenship et Kay, 1964; Cook, 1971). This regularity also observed in our data will allow us to formalise effectively this phenomena[4] and automatically identify them during the preprocess stage.

For us, the preprocessing stage for disfluencies consists in identifying the reparandum and the repair. At the final state of the process, the reparandum would be assigned a specific tag in order to only take into account the repair part in further automatic analyses (e.g. morphosyntactic tagging, chunking, etc.).

2.1. Repetitions

A repetition is a sequence of two (or more) contiguous graphically identical forms. The identical forms can be words or groups of words such in the example below involving repetitions of words *sans* (without) and *la* (the):

ilrMS1 je sais pas / parler sans accent pour moi c'est sans // sans // sans bafouiller sans / sans sans se tromper de mots quoi sans sans sans que la la langue fourche quoi [ilrMS1r]

ilrMS1 I don't know / to speak without an accent for me it's without // without // stamming without / without without getting the wrong word what without without without a slip of the tongue what

Repetitions temporary break the linearity of the statement, by staying on the same location of the syntagmatic axis. The grid representation proposed by (Blanche-Benveniste *et al.*, 1979), allows for taking account of the phenomenom. It superposes repeated terms:

ilrMS1 je sais pas / parler sans accent pour moi
c'est sans
// sans
// sans bafouiller sans
/ sans
sans se tromper de mots quoi sans
sans
sans que la
la langue fourche
quoi

[4] It is based on a systematic linguistic study of the disfluencies and silent pause marks occurring in a 440.000-word corpus (around 40 hours of speech). For more details, see Dister, 2007.

ilrMS1 je sais pas / parler sans accent pour moi c'est **{sans // sans //, .IGN+rep }** **sans** bafouiller **{sans / sans, .IGN+rep }** **sans** se tromper de mots quoi **{sans sans, .IGN+rep }** **sans** que **{la,.IGN+rep}** **la** langue fourche quoi

The reparandum plus interregnum parts are tagged between curly brackets with the tag IGN+rep (IGN for ignore and rep for repetition).

2.2. Immediate self-correction

Immediate self-correction phenomena are variants of repetition ones. In self-correction, one of the morphosyntactic features of the repeated element varies, as it is shown below.

ileFN1 et le journalisme et puis euh **le** les études de journalisme en soi ne me plaisaient pas [ileFN1r]

and I did not like journalism and er studying journalism in itself

In the example, *les* is the plural form of *le*.

The tagging is the same as the one for repetitions, except the tag *cor* for *correction*.

ileFN1 et le journalisme et puis euh **{le,.IGN+cor}** **les** études de journalisme en soi ne me plaisaient pas [ileFN1r]

2.3. Word fragments

A word fragment consists of an interruption of the morpheme being enunciated. According to the terminology in Pallaud (2002), word fragments can be divided into three categories: completed word fragments, corrected word fragments and unachieved word fragments. The three cases are illustrated respectively in the following statements:

iljDV1 apprendre ça c'est transm/ transmettre un savoir donc ça c'est apprendre communiquer euh euh (…) [iljDV1r]

to teach this is to pass a knowledge on so this is to teach to communicate er er

accFJ1 (…) j'ai été à plusieurs reprises avec mes parents en Auvergne je trouvais aussi qu'ils avaient aussi un accent qui était pas mal euh // typique / par contre les J/ les Bretons j'ai jamais su / jamais vu qu'ils avaient d'accent moi (…) [accFJ1r]

I've been on many occasions with my parents in the Auvergne so I found they had too an accent which was quite er // typical / on the other hand / Bretons I've never known / never seen they had an accent I (…)

ilrVI2 m quand un néerlandophone parle français / euh je trouve que ça ne fait pas bien du tout / par rapport à quelqu'un qui parle bien fran/ comme un Bruxellois par exemple (…) [ilrVI2r]

m when a Dutch speaker speaks French / er I find it doesn't make it at all / in comparison with somebody who speaks good Fren/ like a person from Brussels for instance (…)

The two first types of word fragments are subject to the same type of annotation as repetitions and immediate self-correction (the tag is *frag*). The unachieved word fragments are also annotated but without the repair part.

3. Text segmentation

The prepocessing is not only limited to the tagging of the disfluencies. It also requires a new segmentation of the texts, that consists in extracting internal overlapping segments and segmenting speaking slots into smaller parts.

3.1. Speaking slots and overlapping segments

Like disfluencies, overlapping markers break the linearity of the reading. We though observed that in almost all cases, a speech overlapping is not a syntactic break of the statement: the speaker being overlapped continues speaking as if he/she were not interrupted. In the preprocessing stage, we annotate the starting and ending markers of the overlapping segments with the tags IGN+over and.IGN+overEnd. The speaking slots are identified by unique numbers (e.g. #245). For internal overlapping fragments, we extract them in the form of a new speaking slot being referred by the overlapped speech fragment (e.g. @246). Both examples given in section 1.3 are respectively transformed as follows:

{#123,.IGN+slot} {L1,.IGN+speaker} je le connais {|-,.IGN+over} depuis longtemps
{#124,.IGN+slot} {L2,.IGN+speaker} oui tu {-|,.IGN+overEnd} l'avais rencontré à mon mariage

{#245,.IGN+slot} {L1,.IGN+speaker} je l'aime {|-,.IGN+over} vraiment vraiment beaucoup @246{-|,.IGN+overEnd} ce chercheur
{#246,.IGN+slot} {L1,.IGN+speaker}je sais

3.2. Sentence segmentation

Traditionally, Natural Language Processing tools work on the sentence level. Therefore, the first task to do is to segment the graphical chain into tokens (roughly speaking words) and in sentences. Nevertheless, our transcriptions do not contain any punctuation marks and the only a priori segmentation available is the one in speaking slots. As some slots are very long, it is necessary to cut the text into smaller units. For this, we examined whether silent pauses could be the basis of a relevant initial segmentation for automatic annotation. Following studies by Duez (1991) and Candea (2000), we made the hypothesis that transcribers put silent pauses at preferential location in terms of syntactic structure of the text, allowing for relevant regrouping for automatic analysis. In practice, it seems that long pause and silence marks are good candidates for text segmentation in smaller units corresponding roughly to chunks (Abney, 1991), with a low error rate (see Dister, 2008 for further details of the analysis).

4. Outputs

The preprocessing part handles other phenomena in spoken texts such as the speaker identification, the tagging of *euh* (uh), phonetic or paralinguistic markers, etc. The principle is the same: tagging text portions with tag IGN indicating to the analyzer that it has to ignore it. The sample given in section 1.4 is then transformed by our tool in:

{S}{#1,.IGN+slot}{ileGF0,.IGN+speaker} {une,.IGN+rep} une trémie {/,.IGN+meta} ça veut dire quoi {S}
{S}{#2,.IGN+slot} {ilePA2,.IGN+speaker} une trémie justement une trémie {i/,.IGN+frag} {|-@3,.IGN+over} c'est {une,.IGN+corr} {-|,.IGN+overEnd} un tunnel une trémie chez nous {c'est,.IGN+rep} {/,.IGN+meta} c'est le {c'est,.IGN+rep} c'est ce qu'on appelle un tunnel {S}
{S}{#3,.IGN+slot} { ileGF0,.IGN+speaker} oui {S}
{S}{#4,.IGN+slot} {ileGF0,.IGN+speaker} ah d'accord {S}
{S}{#5,.IGN+slot} {ilePA2,.IGN+speaker} hein {|-@6,.IGN+over} mais {-|,.IGN+overEnd} {une pet/,.IGN+frag} un petit tunnel qui n'est pas très long {S}

{S}{#6,.IGN+slot} {ileGF0,.IGN+speaker} mm {S}
{S}{#7,.IGN+slot} {ileGF0,.IGN+speaker} mm {S}
{S}{#8,.IGN+slot} {ilePA2,.IGN+speaker} or une trémie {euh,.IGN+euh} grammaticalement c'est une chose {qui s'en/,.IGN+frag} qui s'enfonce plutôt dans la terre {S}

5. Conclusions

The specificities of texts transcripted from speech rise problems for syntactic and morphosyntactic analyzers. In this paper, we described a preprocessing tool allowing for handling the difficulties inherent to speech transcriptions. The good results obtained by a morphosyntactic tagger (Dister, 2007) and a chunker chunks (Blanc *et al.*, 2008) show the relevancy of our choices.

6. References

ADDA-DECKER, Martine, HABERT, Benoît, BARRAS, Claude, ADDA, Gilles, BOULA DE MAREÜIL, Philippe, PAROUBEK, Patrick (2003). "A disfluency study for cleaning spontaneous speech automatic transcripts and improving speech language models", *ISCA Tutorial and Research Workshop on Disfluency in Spontaneous Speech (DiSS'03)*, Gothenburg, University of Gothenburg, pp. 67-70.

ABNEY, Steve (1991). 'Parsing by chunks', R. Berwick, S. Abney et C. Tenny (eds.), *Principle-based parsing: Computation and Psycholinguistics*, Boston, Kluwer Academic Publishers, pp. 257-278.

BALLY, Charles (1935). *Le Langage et la Vie*, Zurich, Max Niehans (2e éd.).

BÉGUELIN, Marie-José (dir.) (2000). *De la phrase aux énoncés : grammaire scolaire et descriptions linguistiques*, Bruxelles, De Boeck & Larcier.

BÉGUELIN, Marie-José (2002). « Clause, période ou autre ? La phrase graphique et la question des niveaux d'analyse », *Verbum* XXIV 1-2 (*Y a-t-il une syntaxe au-delà de la phrase ?*, M. Charolles, P. Le Goffic et M.-A. Morel Ed.), pp. 85-107.

BÉNARD, Frédérique (2005). *Normalisation de corpus oraux : des métadonnées à l'annotation des transcriptions*, Université Paris-3,

Sorbonne Nouvelle, Mémoire de maitrise.

BENZITOUN, Christophe (2004). « L'annotation syntaxique de corpus oraux constitue-t-elle un problème spécifique ? », *Actes de RÉCITAL* (21 avril 2004, Fès).

BENZITOUN, Christophe, CAMPIONE, Estelle, DEULOFEU, José, HENRY, Sandrine, SABIO, Frédéric, TESTON, Sandra, VALLI, André, VÉRONIS, Jean (2004). « L'analyse syntaxique de l'oral : problèmes et méthode », *Journée d'étude de l'ATALA sur l'annotation syntaxique de corpus* (15 mai 2004, Paris).

BERTHOUD, Anne-Claude, MONDADA, Lorenza (eds) (2000). *Modèles du discours en confrontation*, Berne, Peter Lang.

BLANC, Olivier, DISTER, Anne, CONSTANT, Matthieu et WATRIN, Patrick (2008). « Corpus oraux et chunking », *Actes des 27es Journées d'étude sur la parole (JEP 2008)*, Avignon, 9-13 juin 2008.

BLANCHE-BENVENISTE, Claire, BOREL, Bernard, DEULOFEU, José, DURAND, Jacky, GIACOMI, Alain, LOUFRANI, Claude, MEZIANE, Boudjema, PAZERY, Nelly (1979). « Des grilles pour le français parlé », *Recherches sur le français parlé* 2, Université de Provence, pp. 163-205.

BLANCHE-BENVENISTE, Claire, JEANJEAN, Colette (1987). *Le Français parlé. Transcription et édition*, Paris, Didier Érudition.

BLANCHE-BENVENISTE, Claire, BILGER, Mireille, ROUGET, Christine, VAN DEN EYNDE, Karel (1990). *Le Français parlé. Études grammaticales*, Paris, CNRS Éditions.

BLANKENSHIP, Jane, KAY, Christian (1964). "Hesitation phenomena in English Speech: a study in distribution", *Word* 20, pp. 360-372.

BOOMER, Donald S., DITTMAN, Allen T. (1962). "Hesitation pauses and juncture pauses in speech", *Language and Speech* 5, pp. 215-220.

CANDEA, Maria (2000). *Contribution à l'étude des pauses silencieuses et des phénomènes dits « d'hésitation » en français oral spontané*, Université de Paris-3 Sorbonne-nouvelle, Thèse non publiée.

CHEEPEN, Christine (1995). "Discourse considerations in transcription and analysis", G. Leech, G. Myers, J. Thomas (eds.), *Spoken English on Computer. Transcription, Mark-up and Application*,

New York, Longman, pp. 135-143.

COOK, Mark (1971). "The Incidence of Filled Pauses in Relation to Part of Speech", *Language and Speech* 14, pp. 135-150.

CLÉMENT, Lionel (2001). *Construction et exploitation d'un corpus syntaxiquement annoté pour le français*, Thèse non publiée, Université Paris-7.

COURTOIS, Blandine (1990). « Un système de dictionnaires électroniques pour les mots simples du français », *Langue française* 87, Paris, Larousse, pp. 11-22.

DISTER, Anne (2007). *De la transcription à l'étiquetage morphosyntaxique. Le cas de la banque de données textuelles orales VALIBEL*, Thèse non publiée, Université de Louvain.

DISTER, Anne (2008). « La notation subjective de la pause constitue-t-elle un bon indice pour le découpage de corpus oraux ? », *Description linguistique pour le traitement automatique du français, Cahiers du Cental 5* (M. Constant, A. Dister, L. Emirkanian, S. Piron eds), Louvain-la-Neuve, Presses universitaires de Louvain, pp. 165-186.

DISTER, Anne et SIMON, Anne Catherine (2007). « La transcription synchronisée des corpus oraux. Un aller-retour entre théorie, méthodologie et traitement informatisé. », *Corpus and text linguistics in Romance languages, Arena Romanistica* 1/1, Presses de l'Université de Bergen, Bergen, pp. 54-78.

DISTER, Anne, FRANCARD, Michel, GERON, Geneviève, GIROUL, Vincent, HAMBYE, Philippe, SIMON, Anne Catherine, WILMET, Régine (2006). *Conventions de transcription régissant les corpus de la banque de données VALIBEL*. Available on-line from http://valibel.fltr.ucl.ac.be, corpus oraux, conventions de transcription.

DUEZ, Danielle (1991). *La Pause dans la parole de l'homme politique*, Paris, Éditions du CNRS.

EDWARDS, Jane A. (1995). "Principles and alternative systems in the transcription, coding an mark-up of spoken discourse", *Spoken English on computer. Transcription, mark-up and application* (G. Leech, G. Myers, T. Jenny eds), New York, Longman, pp. 19-34.

FRIBURGER, Nathalie, DISTER, Anne et MAUREL, Denis (2000).

« Améliorer la reconnaissance automatique des fins de phrases », *Actes des troisièmes journées Intex* (A. Dister Éd.), *Revue, Informatique et Statistiques dans les sciences humaines* 36, Université de Liège, pp. 181-199.

FRANCARD, Michel, PERONNET, Louise (1989). « La transcription de corpus oraux dans une perspective comparative. La démarche du projet PLURAL », *Recherche en linguistique appliquée à l'informatique (RELAI)*, CIRB, Québec, pp. 295-307.

GARSIDE, Roger (1995). "Grammatical tagging of the spoken part of the British National Corpus: a progress report", G. Leech, G. Myers, J. Thomas (eds.), *Spoken English on Computer. Transcription, Mark-up and Application*, New York, Longman, pp. 161-167.

GUÉNOT, Marie-Laure (2005). « Parsing de l'oral : traiter les disfluences », *Actes de TALN 2005* (6-10 juin, Dourdan).

LEVELT, Willem J.M. (1989). *Speaking: from intention to articulation*. Cambridge, MIT Press.

MERTENS, Piet (2002). « Les corpus de français parlés ELICOP : consultation et exploitation », J. Binon, P. Desmet, J. Elen, P. Mertens, L. Sercu (eds.), *Tableaux Vivants. Opstellen over taal-en-onderwijs aangeboden aan Mark Debrock*, Leuven, Universitaire Pers.

MONDADA, Lorenza (2000). « Les effets théoriques des pratiques de transcription », *LINX*, 42, revue de l'Université de Paris X-Nanterre, pp. 131-150.

NIVRE, Joakim, GRÖNQVIST, Leif (2001). "Tagging a Corpus of Spoken Swedish", *International Journal of Corpus Linguistics* 6 (1), pp. 47-78.

OCHS, Elinor (1979). "Transcription as theory", *Developmental pragmatics* (E. Ochs et B. B. Schieffelin eds), New York, San Francisco, London, Academic Press, pp. 43-72.

OOSTDIJK, Nelleke (2003). "Normalization and disfluencies in spoken language data", S. Granger et St. Petch-Tyson (eds.), *Extending the scope of corpus-based research. New applications, new challenges*, Amsterdam-New York, Rodopi, pp. 59-70.

PALLAUD, Berthille (2002). « Les amorces de mots comme faits

autonymiques en langage oral », *Recherches sur le français parlé* 17, Université de Provence, pp. 79-101.

PALLAUD, Berthille (2004). « La transgression et la variation », *Marges Linguistiques* 8, pp. 76-87.

PAUMIER, Sébastien (2006). *Unitex 1.2. Manuel d'utilisation*. Available on-line from http://www-igm.univ-mlv.fr/~unitex/manuel.html.

SHRIBERG, Elizabeth (1994). *Preliminaries to a Theory of Speech Disfluencies*, Université de Berkeley, Thèse non publiée.

SILBERZTEIN, Max (1993). *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX*, Paris, Masson.

SIMON, Anne Catherine. (2001). « Le rôle de la prosodie dans le repérage des unités textuelles minimales », *Cahiers de linguistique française* 23, pp 99-125.

VALLI, André, VÉRONIS, Jean (1999). « Étiquetage grammatical des corpus de parole : problèmes et perspectives », *Revue française de linguistique appliquée* 4 (2), pp. 113-133.

VÉRONIS, Jean (2000). « Annotation automatique de corpus : panorama et état de la technique », J.-M. Pierrel (ed.), *Ingénierie des langues*, Paris, Hermès, pp. 111-129.