

Introduction to Corpus Linguistics: Theoretical and methodological basics

Dr. Julien Perrez (ULiège)

7th Young Linguist's Meeting in Poznan

Rethinking language and identity in the multilingual world

23–25 April 2021



1

Comment participer ?



WEB

- 1 Connectez-vous sur www.wooclap.com/YZTJOP
- 2 Vous pouvez participer



SMS

- 1 Pas encore connecté ? Envoyez **@YZTJOP** au **0460 200 711**
- 2 Vous pouvez participer





2

2

Contents

1. Corpus Linguistics: basic concepts
2. Types of corpora
3. Collecting and using corpora



3

References



4

1. Corpus Linguistics: Basic concepts



5

5

1. Corpus Linguistics: Basic concepts

- 1) Introduction
- 2) Theoretical implications
- 3) Concordances
- 4) Collocations
- 5) Frequency
- 6) Corpus annotation



6

6

Introduction to Corpus Linguistics

- Corpus Linguistics (CL) started in the 60's (> computer science)
- Corpus = central notion
 - **General definition:** « a collection of written texts, especially the entire works of a particular author or a body of writing on a particular subject: the Darwinian corpus. » (Oxford OED)
 - **Specific definition:** a collection of written or spoken material in machine-readable form, assembled for the purpose of linguistic research. (Oxford OED)
- « Corpus linguistics is the study of **authentic language data** on a large scale – the **computer aided analysis** of very extensive collections of transcribed utterances or written texts. » (McEnery & Hardie 2012)
- « CL deals with a collection of **machine-readable texts** which is deemed an appropriate basis on which to study a specific set of research questions. » (McEnery & Hardie 2012)



7

7

Machine-readable text

Période 1 (P1)

Animateur (A) : J'espère que personne n'a rien contre l'enregistrement. Ça facilite grandement le travail de Min. Moi c'est <Animateur>, voici <Observateur> qui se présentera aussi, je ne vais pas parler pour lui. <Observateur> est observateur. On sera donc un couple pour la journée. On va essayer de vraiment faciliter les débats, s'assurer que tout le monde parle et puisse exprimer son point de vue. Je prendrais donc jamais position et j'essayerai également que l'ambiance soit bonne au sein du groupe parce qu'on va passer toute la journée ensemble.

Donc, <Observateur>, lui par contre, travaillera un peu plus avec moi, il prendra des notes, c'est pour cela qu'il est très important de mettre en évidence votre petit carton, comme cela, il peut voir, il peut vous noter, vous repérer grâce à ce petit code.

Au niveau des détails pratiques, on va vous demander, si le micro fonctionne, de ne parler que quand vous avez le micro. Cela permet d'abord de faciliter l'enregistrement.

<Test micro>

Donc, voilà, de toute façon, on a cette sauvegarde-là. L'idéal, si on a le micro, c'est de se le passer pour prendre la parole parce que cela permet également de ne pas parler en même temps. C'est assez facile au niveau pratique.

Vous le savez, moi, ici, je suis un ami et un collègue de Min. je fais également une thèse de doctorat ici en sciences politiques. Ce n'est pas du tout sur le fédéralisme. Par contre, le pourquoi vous êtes ici aujourd'hui, c'est le, comme vous le savez sans doute, c'est parce que Min mène sa thèse de doctorat sur le fédéralisme et mesure les perceptions et les préférences des citoyens. Donc, il essaie de comparer, de faire interagir, et cela de deux communautés linguistiques différentes, en Belgique et au Canada. Donc, il y aura quatre journées comme celles-ci, une en Wallonie, une en Flandre, une à Québec et une en Ontario. Donc voilà, la prochaine sera au Canada. Vous êtes donc les premiers à tenter l'expérience. Donc, un tout grand merci pour lui et pour nous d'être là. C'est très gentil de passer toute la journée avec nous pour cela. La méthode qu'on va employer aujourd'hui, cela s'appelle un sondage délibératif, c'est-à-dire que c'est un mélange de ce que l'on appelle quantitatif, donc un questionnaire comme vous l'avez rempli là, ou vous avez les questions comme cela, à choix multiples. On s'est rendu compte en réalité que répondre à des questionnaires comme ceux-là, ce n'est parfois pas la meilleure façon de pouvoir mesurer par exemple les perceptions, ou les préférences d'acteurs par rapport à une problématique particulière. Donc, l'idéal, or ici, d'après Min, l'idéal est la méthodologie qu'il utilise, c'est d'employer, de combiner cette méthode de questionnaire avec des entretiens, donc beaucoup de qualitatif, de permettre aux gens d'interagir de manière différente, enfin, d'interagir en réalité car au niveau du questionnaire, c'est un travail très solitaire.

Donc, vous avez ce questionnaire que vous avez rempli. Vous allez au cours de la journée travailler trois fois, je pense, en petits groupes comme ceci. À chaque fois, la journée sera ponctuée par des rencontres avec des experts, donc vous allez rencontrer des gens qui sont considérés comme tels. Vous avez des gens qui ont travaillé au plus haut niveau de l'État, donc des anciens ministres, vous aurez un journaliste, et puis vous commencerez par un professeur de droit constitutionnel qui vous apportera les premiers éclairages sur ce qu'est le fédéralisme. Donc à chaque fois qu'il y aura des questions un peu plus pointues, moi, je vous renverrai automatiquement vers les experts car il y aura un temps qui vous sera réservé pour leur poser toutes les questions que vous souhaitez.

Donc, un questionnaire au départ. Un travail en petit groupe qui est ponctué d'interactions avec les experts et en fin de journée, et là, ce sera, j'insisterai là-dessus, ce sera important car ça ne sera pas le moment le plus agréable parce que on sera fatigué, il sera tard, que vous remplissiez à nouveau le questionnaire. Et là, le but est de voir si peut-être vos préférences ont changées, si la journée et ce que vous avez pu apprendre, si les débats ont été fructueux ou pas, peu importe, ont changé votre perception, vos perceptions, par rapport au fédéralisme en Belgique. Donc, ce sera important pour Min que vous remplissiez le questionnaire avant de filer.



8

8

Introduction to Corpus Linguistics

- Focus on the analysis of authentic data => **linguistic revolution**
 - « Importantly, the development of corpus linguistics has also spawned, or at least facilitated the exploration of, new theories of language – **theories which draw their inspiration from attested language use** and the findings drawn from it »
- >< Chomsky: **introspection**
 - Chomsky (1984): « if you sit and think for a few minutes, you're just flooded with relevant data »



9

9

Introduction to Corpus Linguistics

be readily available via introspection? Chomsky (1984: 44) sums up the supposed power of introspection by saying that 'if you sit and think for a few minutes, you're just flooded with relevant data'. An example of spontaneous access to, and use of, such data by Chomsky can be seen in the following exchange (Hill, 1962: 29):

Chomsky: The verb *perform* cannot be used with mass word objects: one can *perform a task* but one cannot *perform labour*.

Hatcher: How do you know, if you don't use a corpus and have not studied the verb *perform*?

Chomsky: How do I know? Because I am a native speaker of the English language.



McEnery, T. & A. Hardie (2012). *Corpus Linguistics*. Cambridge : CUP, p. 27

10

10

Introduction to Corpus Linguistics

I have two observations to make. The first is that I don't think there can be any corpora, however large, that contain all of the areas of English lexicon and grammar that I want to explore; all that I have seen are inadequate. The second observation is that every corpus that I've had the chance to examine, however small, has taught me facts that I couldn't imagine finding out in any other way . . .

Fillmore (1992) in McEnery, T. & A. Hardie (2012). *Corpus Linguistics*. Cambridge : CUP, p. 27.



11

11

Introduction to Corpus Linguistics

« Corpus-based Critical Discourse Analysis enables the exploration of patterns of language use which are not observable to the human eye »

Helen Sauntson, YLMP21, 23/04/21



12

12

Introduction to Corpus Linguistics

- Focus on the analysis of authentic data => **linguistic revolution**
 - « Importantly, the development of corpus linguistics has also spawned, or at least facilitated the exploration of, new theories of language – **theories which draw their inspiration from attested language use** and the findings drawn from it »
 - >< Chomsky: **introspection**
 - Chomsky (1984): « if you sit and think for a few minutes, you're just flooded with relevant data »
- CL: Change of perspective in the way linguistic data are considered
- Turns linguistics into an **empirical science**: based on the observation of authentic data
- **Usage-based** grammar

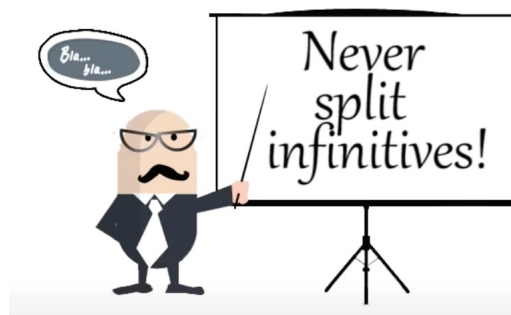


13

13

Split infinitive

- Grammar from usage: Split infinitive
 - « A split infinitive is an English-language grammatical construction in which a word or phrase, usually an **adverb or adverbial phrase**, occurs between the marker **to** and the **bare infinitive (uninflected) form of a verb**. One of the most famous split infinitives occurs in the opening sequence of the Star Trek television series: "**to boldly go where no man has gone before.**" Here, the adverb "boldly" splits the full infinitive "to go."
 - As the split infinitive became more popular in the 19th century, some grammatical authorities sought to introduce a prescriptive rule against it.



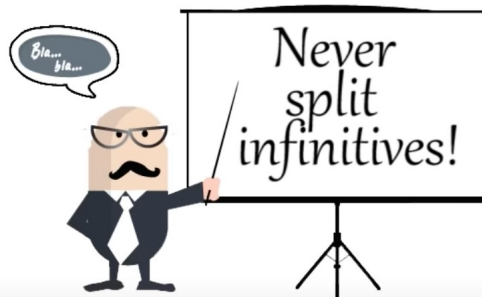
Source: Schools Wikipedia Selection, Loock 2016



14

14

Split infinitive



Split Infinitive Rules

Traditionally, grammar students were always taught not to split their infinitives. The rule dates back as early as the Victorian Era, when Henry Alford advised against splitting infinitives in his 1864 book The Queen's English. Strict grammarians also dislike split infinitives because they interrupt the unit of thought – the infinitive – with a modifier. It separates “to” from its verb, which can make the sentence confusing.

Style guides from the 20th and 21st centuries don't directly speak out against split infinitives. However, they do advise writers against awkward sentence construction that often includes splitting infinitives. The general rule in English is to **avoid splitting infinitives if you can**.

<https://grammar.yourdictionary.com>



15

15

Split infinitive



« You wanna hear the funniest thing ever?

I also split an infinitive and she didn't notice. »



16

16

Split infinitive

Multiple registers :

- (1) a. But in practice Dennis was too repressed to actually go through with it. (BNC W_fict_prose BMR)
- b. What influences them inside their heads to actually go for a certain destination. (BNC S_lect_soc_science F88)
- c. Five of the grandchildren went to Euro Disney on a school trip and loved it but didn't have time to really do it justice. (BNC W_newsp_tabloid CH2)
- d. But just the same, there aren't a lot of people who take it serious enough to actually go looking for the greenies. (COCA FIC FantasySciFi)
- e. It takes some courage to actually go and engage people, but we don't have an alternative to it. (COCA MAG USCatholic)
- f. Simon Johnson, a professor at the Sloan School of Management at M.I.T., and a well-known blogger on banking issues, says he believes that it will take \$1 trillion to really do the trick. (COCA NEWS NYTimes)



From: Loock, R. (2016). *La traductologie de corpus*. Villeneuve d'Ascq: Presses universitaires du Septentrion.

17

17

Introduction to Corpus Linguistics

- CL relies on analysis of **authentic language data**
- « What is corpus linguistics? It is certainly quite distinct from most other topics you might study in linguistics, as it is not directly about the study of any particular aspect of language. **Rather, it is an area which focuses upon a set of procedures, or methods, for studying language** » (McEnery & Hardie 2012: 1)
 - Concordances
 - Collocations (co-occurring patterns)
 - Frequency



18

18

Introduction to Corpus Linguistics

• Concordance

- Corpus analysis tools make it possible to retrieve a sequence of characters of any length in the corpus (suffix, word, multi-word unit, a combination, a sentence)...
- Authentic examples of how this unit is used in a particular language
- Examples: <https://www.english-corpora.org>



19

19

The screenshot shows the COCA interface with the search term 'perform' selected. The concordance results are displayed in a table with columns for line number, search icon, and the text snippet. The word 'perform' is highlighted in green in the original image. A 'Concordance' title is overlaid on the bottom right of the table.

Line	Search Icon	Text Snippet
1	🔍	may be voted off to China when they fail to perform their assigned duties. With ideas like these, and other innovations yet to
2	🔍	Norway, so he'll be heading back there to perform soon. He also is building fan bases in Australia, Dubai and New
3	🔍	free-thinking time," said Hull, who went onto perform and tour with Miles, a disciple of Jimi Hendrix. # NOTE:
4	🔍	Fairfield County? Do you remember seeing a rock legend perform at your school? Please share your
5	🔍	just as the heart and liver have different jobs to perform in the physical body. When humans are able to place themselves in a
6	🔍	to ourselves) was so happy to be able to perform the original show We Are Human, A Story of Standing on Common Ground
7	🔍	proactive on the other. The mental calculus defenders must perform on every read they make is as important to their success as any physical
8	🔍	to contain even whether or not the complex systems will perform the date it is for your victims frighteningly! There are rapidly able tools
9	🔍	tests treatments based on those ideas that they do not perform better than placebo. # All of this explains the propensity of quacks and
10	🔍	Almost nobody chooses to do this style of analysis. Perform the analysis! Kevin Hillstrom, President, MineThatData # Kevin is President
11	🔍	training or races with discretion, allowing the coach to perform the role of " teacher " in those moments. # Explain to your
12	🔍	This means that a couple of octogenarians who could not perform , or had no interest in performing, the marriage act could not marry
13	🔍	current marriage. You are still married. Until you perform the legal requirements that dissolve your legal, civil marriage. # You want
14	🔍	delays the inevitable hard work you're going have to perform if you want to be successful in the long run. If that's
15	🔍	all vertical. # In other words, if you perform a search for cooking, you get a million variations of cooking, but
16	🔍	, that you must often create from scratch, plus perform a myriad of duties that you
17	🔍	sure. However, we all know people who did perform well yet did not get granted tenure (ask me how)
18	🔍	some people who have fantastic academic CVs who can not perform the most basic or workplace tasks, including plain writing and proper use of
19	🔍	around a lot of baggage that impairs their ability to perform some types of work (e.g. it offends personal principles or ideology).
20	🔍	pregnancy, the doctor will usually evaluate her symptoms, perform a pelvic exam and abdominal ultrasound -- the same tests used to feel and

20

ON CLICK: (??) (?)

HELP	?	ALL FORMS (SAMPLE): 100 200 500	FREQ	TOTAL UNIQU
1	<input type="checkbox"/>	CRITICISM	32196	
2	<input type="checkbox"/>	TERRORISM	25593	
3	<input type="checkbox"/>	RACISM	18658	
4	<input type="checkbox"/>	MECHANISM	17801	
5	<input type="checkbox"/>	JOURNALISM	16368	
6	<input type="checkbox"/>	CAPITALISM	13778	
7	<input type="checkbox"/>	AUTISM	10432	
8	<input type="checkbox"/>	TOURISM	9863	
9	<input type="checkbox"/>	COMMUNISM	8667	
10	<input type="checkbox"/>	OPTIMISM	8455	
11	<input type="checkbox"/>	SOCIALISM	8325	
12	<input type="checkbox"/>	NATIONALISM	7106	
13	<input type="checkbox"/>	FEMINISM	6565	
14	<input type="checkbox"/>	SKEPTICISM	6449	
15	<input type="checkbox"/>	ACTIVISM	5954	
16	<input type="checkbox"/>	REALISM	5445	
17	<input type="checkbox"/>	LIBERALISM	5033	
18	<input type="checkbox"/>	ORGANISM	4944	
19	<input type="checkbox"/>	CONSERVATISM	4877	
20	<input type="checkbox"/>	METABOLISM	4734	

Concordance

21

ON CLICK: (??) (?)

HELP	?	ALL FORMS (SAMPLE): 100 200 500	FREQ	TOTAL 4 UNIQU
1	<input type="checkbox"/>	PERFORM ABORTIONS	225	
2	<input type="checkbox"/>	PERFORM TASKS	170	
3	<input type="checkbox"/>	PERFORM SURGERY	104	
4	<input type="checkbox"/>	PERFORM MIRACLES	100	
5	<input type="checkbox"/>	PERFORM MUSIC	86	
6	<input type="checkbox"/>	PERFORM LIST	71	
7	<input type="checkbox"/>	PERFORM WORK	65	
8	<input type="checkbox"/>	PERFORM SERVICES	57	
9	<input type="checkbox"/>	PERFORM ACTS	53	
10	<input type="checkbox"/>	PERFORM ACTIONS	52	
11	<input type="checkbox"/>	PERFORM COMMUNITY	50	
12	<input type="checkbox"/>	PERFORM EXPERIMENTS	44	
13	<input type="checkbox"/>	PERFORM FUNCTIONS	44	
14	<input type="checkbox"/>	PERFORM WORKS	42	
15	<input type="checkbox"/>	PERFORM OPERATIONS	41	
16	<input type="checkbox"/>	PERFORM SEX	41	
17	<input type="checkbox"/>	PERFORM DUTIES	36	
18	<input type="checkbox"/>	PERFORM SONGS	36	
19	<input type="checkbox"/>	PERFORM RESEARCH	33	
20	<input type="checkbox"/>	PERFORM MAINTENANCE	32	

Collocations

22

Corpus of Contemporary American English

SEARCH FREQUENCY CONTEXT

EXPLORE NEW FEATURES

produce tangible goods, like clocks or photographs, or **perform labor** or physical services on your physical assets, like your car or house

of " foreigners and aliens under contract or agreement to **perform labor** in the United States "), while applying to an alien brought

alien or aliens, foreigner or foreigners, to **perform labor** or service of any kind in the United States, its territories,

migration of foreigners and aliens under contract or agreement to **perform labor** in the United States, its territories, and the District of Columbia

from bringing foreigners into the United States under contract to **perform labor**. The only exceptions are those immigrants brought to perform domestic service and

. Any person directed, allowed, or permitted to **perform labor** or service of any kind by an employer. The employees of an

from helicopters. # The government hires some workers to **perform labor** intensive work. The govt. instantaneously taxes \$X from the people and transfers

raise money. Those who could not pay had to **perform labor** at the main government stations. For some the taxes forced them into

), in addition, members of households had to **perform labor** services for the soldier-settlers. Failure to pay tribute and to perform labor

for the soldier-settlers. Failure to pay tribute and to **perform labor** service re (women and children)

Collocations

23

Introduction to Corpus Linguistics

- « What is corpus linguistics? It is certainly quite distinct from most other topics you might study in linguistics, as it is not directly about the study of any particular aspect of language. **Rather, it is an area which focuses upon a set of procedures, or methods, for studying language** » (McEnery & Hardie 2012: 1)
 - Concordances
 - Collocations (co-occurring patterns)
 - **Frequency**

24

ON CLICK: (??) (?)

HELP	?	ALL FORMS (SAMPLE): 100 200 500	FREQ	TOTAL UNIQU
1	<input type="checkbox"/>	CRITICISM	32196	
2	<input type="checkbox"/>	TERRORISM	25593	
3	<input type="checkbox"/>	RACISM	18658	
4	<input type="checkbox"/>	MECHANISM	17801	
5	<input type="checkbox"/>	JOURNALISM	16368	
6	<input type="checkbox"/>	CAPITALISM	13778	
7	<input type="checkbox"/>	AUTISM	10432	
8	<input type="checkbox"/>	TOURISM	9863	
9	<input type="checkbox"/>	COMMUNISM	8667	
10	<input type="checkbox"/>	OPTIMISM	8455	
11	<input type="checkbox"/>	SOCIALISM	8325	
12	<input type="checkbox"/>	NATIONALISM	7106	
13	<input type="checkbox"/>	FEMINISM	6565	
14	<input type="checkbox"/>	SKEPTICISM	6449	
15	<input type="checkbox"/>	ACTIVISM	5954	
16	<input type="checkbox"/>	REALISM	5445	
17	<input type="checkbox"/>	LIBERALISM	5033	
18	<input type="checkbox"/>	ORGANISM	4944	
19	<input type="checkbox"/>	CONSERVATISM	4877	
20	<input type="checkbox"/>	METABOLISM	4734	

Frequency

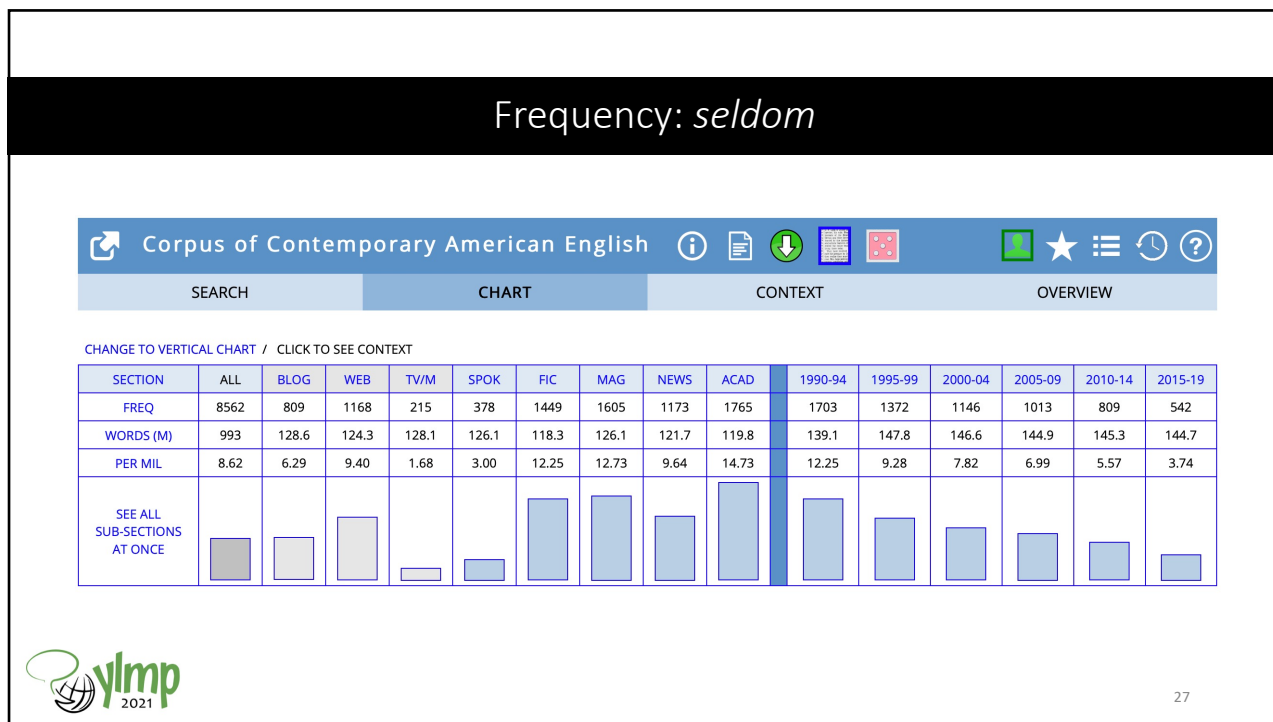
25

ON CLICK: (??) (?)

HELP	?	ALL FORMS (SAMPLE): 100 200 500	FREQ	TOTAL 4 UNIQUE
1	<input type="checkbox"/>	PERFORM ABORTIONS	225	
2	<input type="checkbox"/>	PERFORM TASKS	170	
3	<input type="checkbox"/>	PERFORM SURGERY	104	
4	<input type="checkbox"/>	PERFORM MIRACLES	100	
5	<input type="checkbox"/>	PERFORM MUSIC	86	
6	<input type="checkbox"/>	PERFORM LIST	71	
7	<input type="checkbox"/>	PERFORM WORK	65	
8	<input type="checkbox"/>	PERFORM SERVICES	57	
9	<input type="checkbox"/>	PERFORM ACTS	53	
10	<input type="checkbox"/>	PERFORM ACTIONS	52	
11	<input type="checkbox"/>	PERFORM COMMUNITY	50	
12	<input type="checkbox"/>	PERFORM EXPERIMENTS	44	
13	<input type="checkbox"/>	PERFORM FUNCTIONS	44	
14	<input type="checkbox"/>	PERFORM WORKS	42	
15	<input type="checkbox"/>	PERFORM OPERATIONS	41	
16	<input type="checkbox"/>	PERFORM SEX	41	
17	<input type="checkbox"/>	PERFORM DUTIES	36	
18	<input type="checkbox"/>	PERFORM SONGS	36	
19	<input type="checkbox"/>	PERFORM RESEARCH	33	
20	<input type="checkbox"/>	PERFORM MAINTENANCE	32	

Frequency

26



27

1. Corpus Linguistics: basic concepts

- 1) Introduction
- 2) Theoretical implications
- 3) Concordances
- 4) Collocations
- 5) Frequency
- 6) **Corpus annotation**

28

Corpus annotation

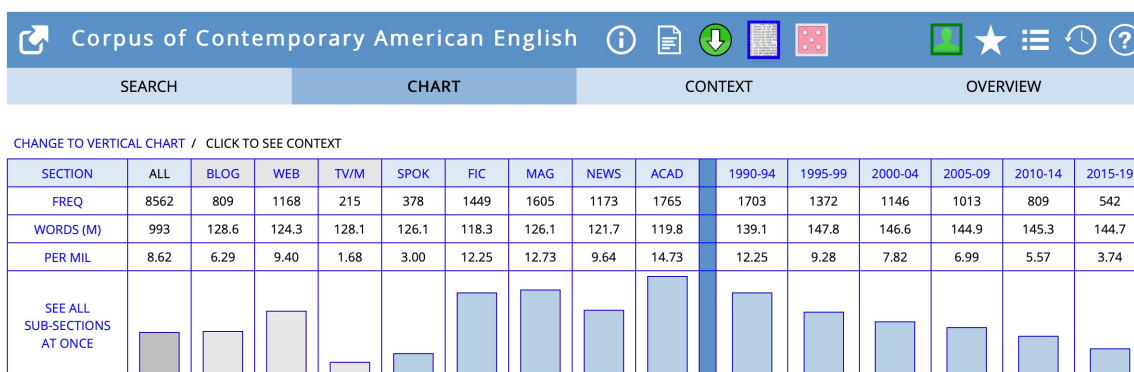
• Metadata

- Information about the text itself – for example, in the case of written material, the metadata may tell you who wrote it, when it was published, and what language it is written in.
- Raw corpus >< textual database
 - Makes it possible to compose subcorpora which can be compared



29

29



30

30

Corpus annotation

- « We can also encode linguistic information within a corpus text in such a way that we can systematically and accurately recover that analysis later ; when this is done, the corpus is said to be *analytically* or *linguistically annotated*. » (McEnery & Hardie 2012)
- **Annotated corpus**
 - Enriched with (linguistic) information about the form or function of units (strings of characters)
 - Automatic vs. manual annotation
- **General** forms of annotation
 - Lemmatization
 - POS-tagging
 - Parsing (syntactic tagging)
 - Semantic tagging
- **Specific** forms of annotation
 - Research specific
 - VU Corpus Metaphor



31

31

Corpus annotation

- **Word form (token)**
 - Researcher, researchers,
 - is, was, be, been
- **Lemma (type)**
 - <RESEARCHER>, <BE>
 - Includes all the inflected forms of a noun or verb



32

32

Corpus annotation > Part-of-Speech tagging

President Biden rejoined the Paris agreement on his first day in office and pledged to hold a leaders summit shortly after

Among those attending will be China's President Xi Jinping.

Despite serious tensions between the two countries on a host of issues, both sides seem keen to keep climate change separate from these disputes. Last weekend, the two countries issued a joint statement saying they would tackle climate "with the seriousness and urgency it demands".

President_NPO Biden_NPO rejoined_VVD the_ATO Paris_NPO agreement_NN1 on_PRP

his_DPS first_ORD day_NN1 in_PRP office_NN1 and_CJC pledged_VVD to_TOO hold_VVI a_ATO leaders_NN2 summit_NN1 shortly_AV0 after_CJS Among_PRP

those_DTO attending_VVG will_VM0 be_VBI China_NPO 's_POS President_NPO Xi_NPO

Jinping_NPO .SENT -----PUN

Despite_PRP serious_AJ0 tensions_NN2 between_PRP the_ATO two_CRD countries_NN2

on_PRP a_ATO host_NN1 of_PRF issues_NN2 ,_PUN both_DTO sides_NN2 seem_VVB

keen_AJ0 to_TO0 keep_VVI climate_NN1 change_NN1 separate_VVB from_PRP

these_DTO disputes_NN2 .SENT -----PUN

Last_ORD weekend_NN1 ,_PUN the_ATO two_CRD countries_NN2 issued_VVD a_ATO

joint_AJ0 statement_NN1 saying_VVG they_PNP would_VM0 tackle_VVI climate_NN1

"_PUQ with_PRP the_ATO seriousness_NN1 and_CJC urgency_NN1 it_PNP demands_VVZ

"_PUQ .SENT -----PUN



33

33

Corpus annotation

Climate change: Biden summit to push for immediate action

By Matt McGrath
Environment correspondent

7 hours ago



34

34

Corpus annotation > Part-of-Speech tagging

- CLAWS part-of-speech tagger for English (University of Lancaster)
- <http://ucrel.lancs.ac.uk/claws/>
- Tags



35

35

HELP	?	ALL FORMS (SAMPLE): 100 200 500	FREQ	TOTAL 4 UNIQUE
1	<input type="checkbox"/>	PERFORM ABORTIONS	225	
2	<input type="checkbox"/>	PERFORM TASKS	170	
3	<input type="checkbox"/>	PERFORM SURGERY	104	
4	<input type="checkbox"/>	PERFORM MIRACLES	100	
5	<input type="checkbox"/>	PERFORM MUSIC	86	
6	<input type="checkbox"/>	PERFORM LIST	71	
7	<input type="checkbox"/>	PERFORM WORK	65	
8	<input type="checkbox"/>	PERFORM SERVICES	57	
9	<input type="checkbox"/>	PERFORM ACTS	53	
10	<input type="checkbox"/>	PERFORM ACTIONS	52	
11	<input type="checkbox"/>	PERFORM COMMUNITY	50	
12	<input type="checkbox"/>	PERFORM EXPERIMENTS	44	
13	<input type="checkbox"/>	PERFORM FUNCTIONS	44	
14	<input type="checkbox"/>	PERFORM WORKS	42	
15	<input type="checkbox"/>	PERFORM OPERATIONS	41	
16	<input type="checkbox"/>	PERFORM SEX	41	
17	<input type="checkbox"/>	PERFORM DUTIES	36	
18	<input type="checkbox"/>	PERFORM SONGS	36	
19	<input type="checkbox"/>	PERFORM RESEARCH	33	
20	<input type="checkbox"/>	PERFORM MAINTENANCE	32	

POS-Tagging: perform + **NOUN**

36

Corpus annotation > Parsing

- Syntactic annotation
- <https://nlp.stanford.edu/software/lex-parser.shtml#Sample>



37

37

Semantic annotation

President Biden rejoined the Paris agreement on his first day in office and pledged to hold a leaders summit shortly after

Among those attending will be China's President Xi Jinping.

Despite serious tensions between the two countries on a host of issues, both sides seem keen to keep climate change separate from these disputes. Last weekend, the two countries issued a joint statement saying they would tackle climate "with the seriousness and urgency it demands".

President_G1.1/S7.1+/S2mf Biden_Z99 rejoined_S5+/N6+ the_Z5 Paris_Z2 agreement_A6.1+ on_Z5 his_Z8m first_N4 day_T1.3 in_Z5 office_I2.1/H1c and_Z5

pledged_A7+/Q2.2 to_Z5 hold_M2 a_Z5 leaders_S7.1+/S2mf summit_W3 shortly_N3.8+

after_Z5 Among_Z5 those_Z8 attending_S1.1.3+ will_T1.1.3 be_A3+ China_Z2 's_Z5

President_G1.1/S7.1+/S2mf Xi_Z1mf[i11.2.1 Jinping_Z1mf[i11.2.2 ._PUNC

Despite_Z5 serious_A11.1+ tensions_E6- between_Z5 the_Z5 two_N1 countries_M7

on_Z5 a_Z5 host_S2mf of_Z5 issues_X4.1 ,_PUNC both_N5.1+[i12.2.1

sides_N5.1+[i12.2.2 seem_A8 keen_X5.2+ to_Z5 keep_A9+ climate_W4 change_A2.1+

separate_A6.1- from_Z5 these_Z5 disputes_A6.1- ._PUNC

Last_T1.1.1[i13.2.1 weekend_T1.1.1[i13.2.2 ,_PUNC the_Z5 two_N1 countries_M7

issued_A9- a_Z5 joint_S5+ statement_Q2.1 saying_Q2.1 they_Z8mfn would_A7+

tackle_A1.1.1 climate_W4 " _PUNC with_Z5 the_Z5 seriousness_A11.1+ and_Z5

urgency_A11.1+ it_Z8 demands_Q2.2 " _PUNC ._PUNC

38



38

Semantic annotation

- <http://ucrel.lancs.ac.uk/usas/>

UCREL Semantic Analysis System (USAS)

[USAS Home Page](#) | [English tagger](#) | [Dutch tagger](#) | [Chinese tagger](#) | [Italian tagger](#) | [Portuguese tagger](#) | [Spanish tagger](#) | [GUI download](#)

English Semantic Tagger

The UCREL semantic analysis system is a framework for undertaking the automatic semantic analysis of text. The framework has been designed and used across a number of research projects and this page collects together various pointers to those projects and publications produced since 1990.

The semantic tagset used by USAS was originally loosely based on Tom McArthur's Longman Lexicon of Contemporary English (McArthur, 1981). It has a multi-tier structure with 21 major discourse fields (shown here on the right), subdivided, and with the possibility of further fine-grained subdivision in certain cases. We have written an [introduction to the USAS category system](#) (PDF file) with examples of prototypical words and multi-word units in each semantic field.

The full tagset is available on-line in [plain text form](#) and [formatted on one page in PDF](#). We also have a list of the [full descriptive labels of the semantic subcategories](#). You can also see the USAS semantic tagset in Russian as a [two page PDF](#) and [text file](#). For other languages, see below.

A [visual representation](#) showing the USAS tagset heirarchy is now on-line, along with those for the [Louw-Nida model](#) and the [Hallig/Von Wartburg/Schmidt/Wilson Model](#).

You can try out the [English semantic tagger online](#).

A general and abstract terms	B the body and the individual	C arts and crafts	E emotion
F food and farming	G government and public	H architecture, housing and the home	I money and commerce in industry
K entertainment, sports and games	L life and living things	M movement, location, travel and transport	N numbers and measurement
O substances, materials, objects and equipment	P education	Q language and communication	S social actions, states and processes
T Time	W world and environment	X psychological actions, states and processes	V science and technology
Z names and grammar			

39

Corpus annotation

- Raw corpus >> **annotated corpus**
 - Enriched with (linguistic) information about the form or function of units (strings of characters)
 - Automatic vs. manual annotation
- **General** forms of annotation
 - Lemmatization
 - POS-tagging
 - Parsing (syntactic tagging)
 - Semantic tagging
- **Specific** forms of annotation
 - Research specific
 - [VU Amsterdam Metaphor Corpus Online](#)



40

40

1. Corpus Linguistics: Basic concepts

- 1) Introduction
- 2) Theoretical implications
- 3) Concordances
- 4) Collocations
- 5) Frequency
- 6) Corpus annotation

Questions?



41

41

2. Types of corpora



42

42

Monolingual corpora

- Reference corpora > Aiming at collecting a representative sample of a language
 - Brown Corpus (American English) – 1,000,000 words – 15 textual genres
 - British National Corpus (BNC): <https://www.english-corpora.org/bnc/>
 - Corpus of Contemporary American English (COCA): <https://www.english-corpora.org/coca/>
 - IPI PAN Corpus for Polish: <http://korpus.pl/>
 - The Russian National Corpus: <http://ruscorpora.ru/>
 - Hungarian National Corpus: http://corpus.nytud.hu/mnsz/index_eng.html
 - Czech National Corpus: <https://www.clarin.eu/lrtshowcases/czech-national-corpus>
 - Corpus of the Contemporary Lithuanian Language (<https://clarin.vdu.lt/xmlui/handle/20.500.11821/16>)
 - The UCLA Written Chinese Corpus: <https://www.lancaster.ac.uk/fass/projects/corpus/UCLA/>
 - Das Deutsche Referenzkorpus – DeReKo: <https://www1.ids-mannheim.de/kl/projekte/korpora/>
 - ...



43

43

Monolingual corpora

- Specialized corpora
 - Focus on a specific genre
 - Soap operas
 - Films
 - Coronavirus
 - ...
 - Focus on a specific research question
 - Learner corpora
 - [International Corpus of Learner English](#) (UCLouvain)
 - Political corpora



44

44



The most widely used online corpora: [guided tour](#), [overview](#), [search types](#), [variation](#), [virtual corpora](#), [corpus-based resources](#), [BYU](#).

The links below are for the online interface. But you can also  download the corpora for use on your own computer.

Corpus (online access)	Download	# words	Dialect	Time period	Genre(s)
iWeb: The Intelligent Web-based Corpus		14 billion	6 countries	2017	Web
News on the Web (NOW)		12.4 billion+	20 countries	2010-yesterday	Web: News
Global Web-Based English (GloWbE)		1.9 billion	20 countries	2012-13	Web (incl blogs)
Wikipedia Corpus		1.9 billion	(Various)	2014	Wikipedia
Corpus of Contemporary American English (COCA)		1.0 billion	American	1990-2019	Balanced
Coronavirus Corpus		974 million+	20 countries	Jan 2020-yesterday	Web: News
Corpus of Historical American English (COHA)		475 million	American	1820-2019	Balanced
The TV Corpus		325 million	6 countries	1950-2018	TV shows
The Movie Corpus		200 million	6 countries	1930-2018	Movies
Corpus of American Soap Operas		100 million	American	2001-2012	TV shows

45

Collecting corpora > balance and representativeness

- « the corpus data we select to explore a research question must be well matched to that research question » (McEnery & Hardie 2012)
- **Reference corpora**
 - Balance between the different genres

46

Table 1.1 *The LOB Corpus Sampling Frame (after Hofland and Johansson 1982: 2)*

Category mnemonic	Description	Number of text samples in this category
A	Press: reportage	44
B	Press: editorial	27
C	Press: reviews	17
D	Religion	17
E	Skills, trades and hobbies	38
F	Popular lore	44
G	Belles lettres, biography, essays	77
H	Miscellaneous (government documents, foundation reports, industry reports, college, catalogue, industry house organ)	30
J	Learned and scientific writings	80
K	General fiction	29
L	Mystery and detective fiction	24
M	Science fiction	6
N	Adventure and western fiction	29
P	Romance and love story	29
R	Humour	9
Total		500



47

47

Corpus of Contemporary American English									
SEARCH	FREQUENCY				CONTEXT			TEXTS	
YEAR	BLOG	WEB	TV / MOVIES	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC	TOTAL
TOTAL	125,496,215	129,899,426	128,013,334	127,396,916	119,505,292	127,352,014	122,959,393	120,988,348	1,001,610,938
1990			3,207,900	4,374,469	4,162,242	4,101,447	4,082,931	3,983,143	23,914,122
1991			3,379,151	4,316,898	4,192,646	4,209,838	4,104,806	4,051,046	24,256,376
1992			3,183,858	4,523,054	3,893,956	4,288,694	4,092,031	4,028,147	24,011,732
1993			3,785,924	4,487,978	3,921,244	4,254,351	4,153,070	4,150,671	24,755,231
1994			4,375,338	4,457,726	3,870,757	4,310,375	4,147,947	4,047,115	25,211,252
1995			5,006,966	4,548,602	3,846,412	4,314,737	4,122,703	4,016,371	25,857,786
1996			4,384,976	4,095,266	3,758,787	4,338,766	4,099,305	4,110,209	24,789,305
1997			4,380,670	3,904,996	3,617,741	4,368,917	4,153,906	4,420,786	24,849,013
1998			4,390,197	4,446,217	3,779,801	4,393,835	4,122,295	4,111,453	25,245,796
1999			4,381,144	4,445,564	4,154,537	4,391,146	4,107,423	4,023,282	25,505,095
2000			4,385,593	4,455,815	3,942,474	4,387,935	4,037,086	4,093,991	25,304,894
2001			4,389,164	4,026,240	3,894,789	4,298,636	4,072,447	3,965,654	24,648,931
2002			4,384,475	4,372,290	3,766,673	4,310,634	4,114,280	4,054,359	25,004,713

48

Collecting corpora > balance and representativeness

- « the corpus data we select to explore a research question must be well matched to that research question » (McEnery & Hardie 2012)
- **Reference corpora**
 - Balance between the different genres
- **Specialized corpora**
 - Internal validity



49

49

Multilingual (parallel) corpora

- Source texts in one language plus translations into one or more other languages
- Pairs or groups of monolingual corpora designed using the same sampling frame
- Examples:
 - Linguee
 - Dutch Parallel Corpus
 - EUR-Lex website (<http://eur-lex.europa.eu>)



50

50

Linguee

[...] vous-même évaluateur, mais vous ne serez jamais invité à évaluer une proposition pour laquelle vous auriez un conflit d'intérêt . <small>↳ cordis.europa.eu</small>	[...] evaluator yourself, although you will never be invited to evaluate a proposal in which you have a conflict of interest . <small>↳ cordis.europa.eu</small>
Les membres du Comité sont invités à s'abstenir de débattre de sujets qui suscitent chez eux un conflit d'intérêt . <small>↳ eur-lex.europa.eu</small>	Members of the Committee should abstain from discussions on a topic on which they have a conflict of interest . <small>↳ eur-lex.europa.eu</small>
Le membre en conflit d'intérêt s'abstiendra de la discussion et du vote portant sur le sujet en cause. <small>↳ kativik.qc.ca</small>	The member in a conflict of interest shall abstain from the discussion and vote on the matter at issue. <small>↳ kativik.qc.ca</small>
[...] impliqué doit fournir, par écrit, un avis détaillé du conflit d'intérêt réel ou perçu, à l'exécutif de la filiale; l'exécutif [...]. <small>↳ legion.ca</small>	In these instances the person involved must provide a detailed, written notice of the real or potential, conflict of interest to the branch executive . <small>↳ legion.ca</small>
Il est important que l'apparence même d'un conflit d'intérêt soit évitée. <small>↳ dorel.com</small>	It is important that even the appearance of a conflict of interest be avoided. <small>↳ dorel.com</small>
Quelques-uns les critiques dénoncent que c'est un conflit d'intérêt , comme mettre le renard pour garder le poulailler. <small>↳ minelinks.com</small>	Some critics charge that this is a conflict of interest , like putting the fox in charge of guarding the hen house. <small>↳ minelinks.com</small>
En cas d'éventuel conflit d'intérêt avec un grand actionnaire de la société, la procédure de l'article 524 du Code des Sociétés est appliqué. <small>↳ investretail.be</small>	In case of a possible conflict of interest with a majority shareholder of the company, the procedure described in article [...] <small>↳ investretail.be</small>



51

51

3. Collecting and using corpora



52

52

3. Collecting and using corpora

- 1) Web as a Corpus
- 2) Web for Corpus
- 3) Collecting your own corpus
- 4) Using online corpora



53

53

1. Web as a corpus

- Gradual shift towards big data corpora (iWeb)
- Google
- [Webcorp](#)
 - « WebCorp is a suite of tools which allows access to the World Wide Web as a corpus - a large collection of texts from which facts about the language can be extracted. » (Webcorp Live)
- [Google Book Corpus](#)
 - Digitized books (more than 5,000,000 books, more than 5,000,000,000 words)
 - Different languages



54

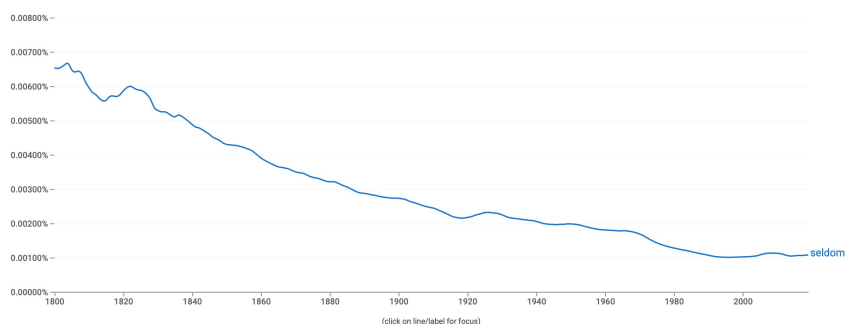
54

Google Books Corpus

Google Books Ngram Viewer

Q seldom X ?

1800 - 2019 English (2019) Case-Insensitive Smoothing



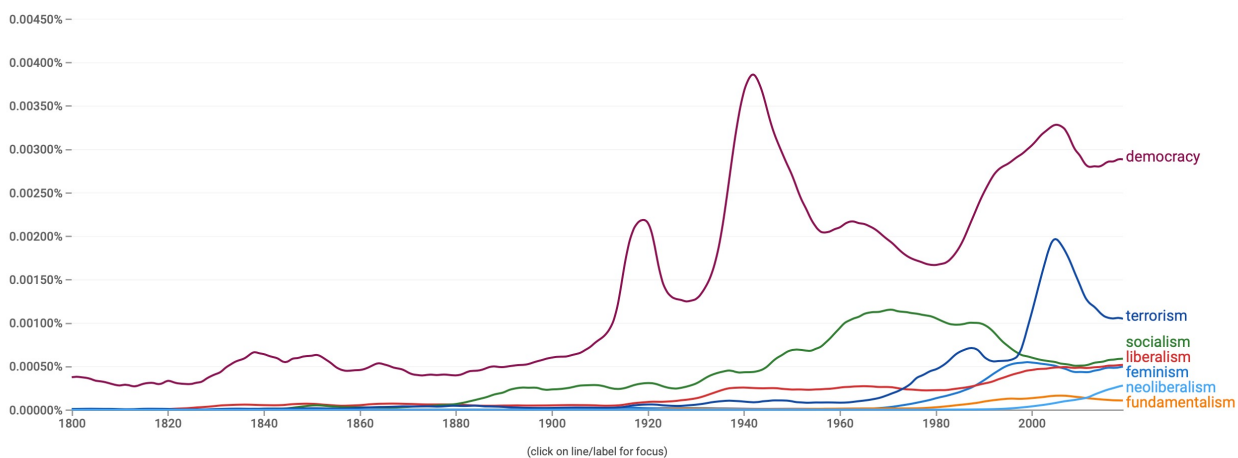
55

55

Google Books Ngram Viewer

Q feminism,liberalism,socialism,fundamentalism,terrorism,democracy,neoliberalism X ?

1800 - 2019 English (2019) Case-Insensitive Smoothing



56

Web as a corpus

- Discussion?



57

57

2. Web for corpus

- Extract data from the web to collect a specific corpus (that can be used offline)
- [Google Custom Search Engine](#)
 - Personal search engine
- [Bootcat](#) (software)



58

58

3. Collecting your own corpus

- Manual collection
- Specific research projects
 - Classroom interactions (H. Sauntson)
 - Political corpora (debates, citizen discourse...)
 - Learner corpora (essays, interactions...)



59

59

4. Using online corpora

- English-Corpora.org
- Sketchengine



66

66

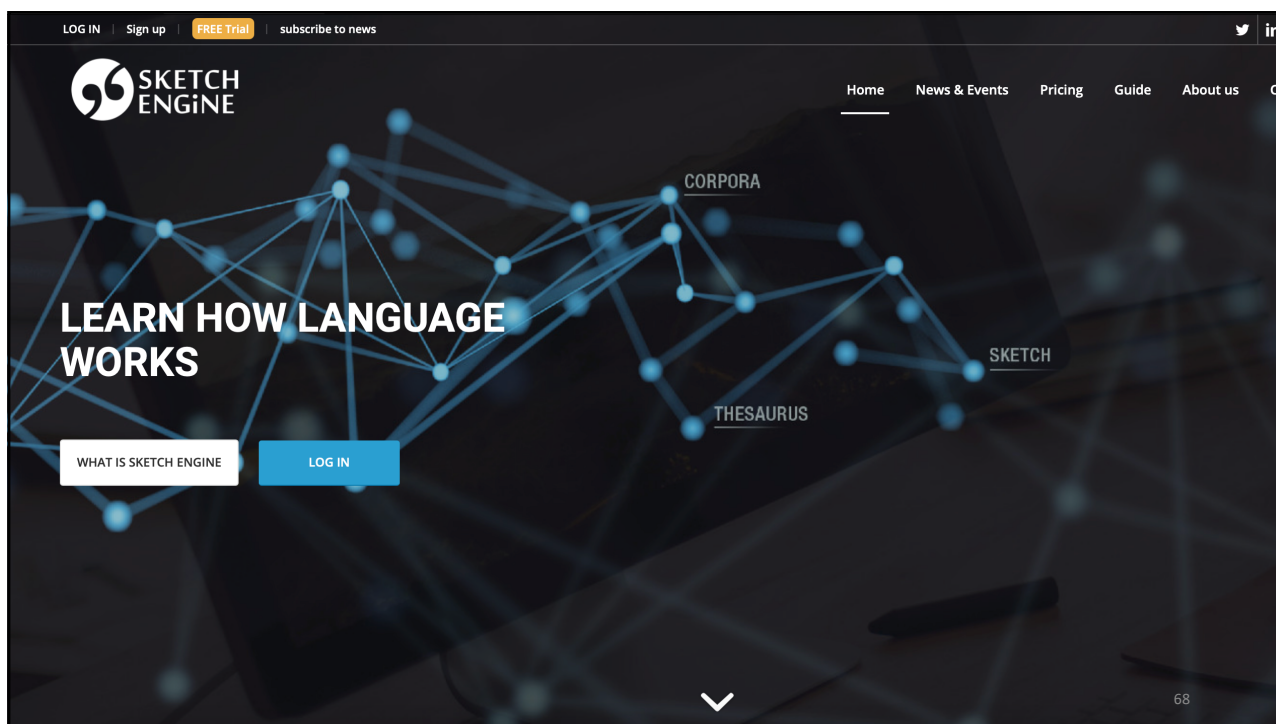


The most widely used online corpora: [guided tour](#), [overview](#), [search types](#), [variation](#), [virtual corpora](#), [corpus-based resources](#), [BYU](#).

The links below are for the online interface. But you can also  download the corpora for use on your own computer.

Corpus (online access)	Download	# words	Dialect	Time period	Genre(s)
iWeb: The Intelligent Web-based Corpus		14 billion	6 countries	2017	Web
News on the Web (NOW)		12.4 billion+	20 countries	2010-yesterday	Web: News
Global Web-Based English (GloWbE)		1.9 billion	20 countries	2012-13	Web (incl blogs)
Wikipedia Corpus		1.9 billion	(Various)	2014	Wikipedia
Corpus of Contemporary American English (COCA)		1.0 billion	American	1990-2019	Balanced
Coronavirus Corpus		973 million+	20 countries	Jan 2020-yesterday	Web: News
Corpus of Historical American English (COHA)		475 million	American	1820-2019	Balanced
The TV Corpus		325 million	6 countries	1950-2018	TV shows
The Movie Corpus		200 million	6 countries	1930-2018	Movies ⁶⁷

67



LOG IN | Sign up | **FREE Trial** | subscribe to news

SKETCH ENGINE

Home | News & Events | Pricing | Guide | About us | Co

CORPORA

LEARN HOW LANGUAGE WORKS

WHAT IS SKETCH ENGINE | LOG IN

SKETCH

THESAURUS

68

68

69

List of references

Anthony, L. (2020). *AntConc (Version 3.5.9)* [Computer Software]. Available from <https://www.laurenceanthony.net/software>.

Reference Guide for the British National Corpus (XML Edition) edited by Lou Burnard, February 2007. URL: <http://www.natcorp.ox.ac.uk/XMLedition/URG/>

Davies, M. (2020, November). *English-Corpora.org: a guided tour*. English-Corpora.org. <https://www.english-corpora.org/pdf/english-corpora.pdf>

Loock, R. (2016). *La traductologie de corpus*. Villeneuve d'Ascq: Presses universitaires du Septentrion.

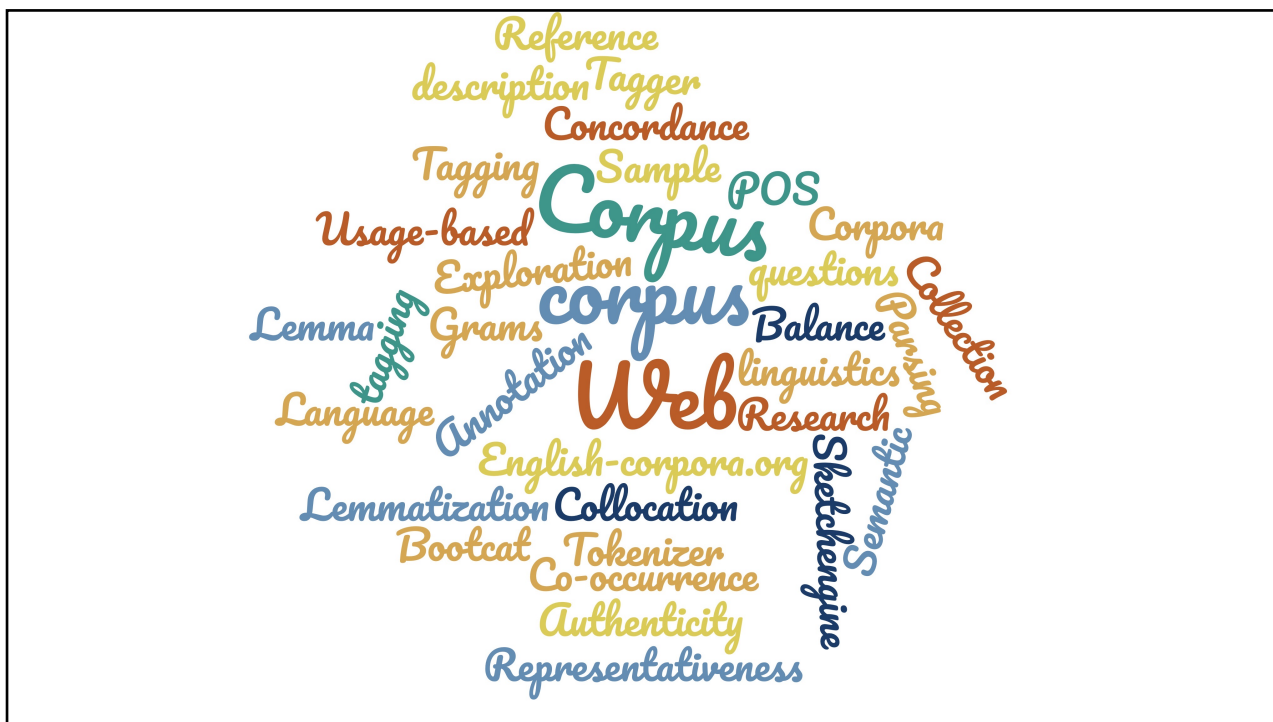
McEnery, T., & Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge. Cambridge University Press.

McEnery, T., & Wilson, A. (2001). *Corpus Linguistics: An Introduction* [2nd edition]. Edinburgh: Edinburgh University Press.



70

70



71