Chiara Messina

Swiss Federal Chancellery

Central Language Services

Terminology Section

Bern, 2018-06-13

Using corpora in terminography[1]

1. Some advantages of using corpora in terminology tasks

A corpus depicts how language is used by speakers (or writers) in daily communication, when talking or writing about a particular subject, or in a specific situation. While general corpora are not very interesting for terminologists, specialised corpora are a useful tool. They deliver a lot of relevant information (for example, contexts that can be used in terminological entries and in most cases data for explicatory notes as well), may back terminologists' assumptions on new terms and their use or, on the contrary, refute their hypotheses. They also help terminologists understand whether a term is still being used or not, and whether it has been replaced by another one.

With the help of specific tools it is possible to extract statistical information from corpora, such as the frequency with which a term is used, which words a term usually combines with, the keywords of a text, and so on. Frequency of use and distribution (patterns) of linguistic units provide information about the status of a term and give clues on concept building and stability, that is, whether the concept is already established or if it is still in a fuzzy phase. "The relevance of corpus-based studies for terminology can be simply illustrated by reference to neologisms and archaisms […] and compound terms, the staple of any terminology enterprise (accessible, for instance, through concordance searches, statistical texts of collocational patterns, or the identification of typical patterns characterizing possible forms)" (Ahmad/Rogers 2001:730). In the case of compound lexical units, corpora may help to understand whether they work as terms in a specific field.

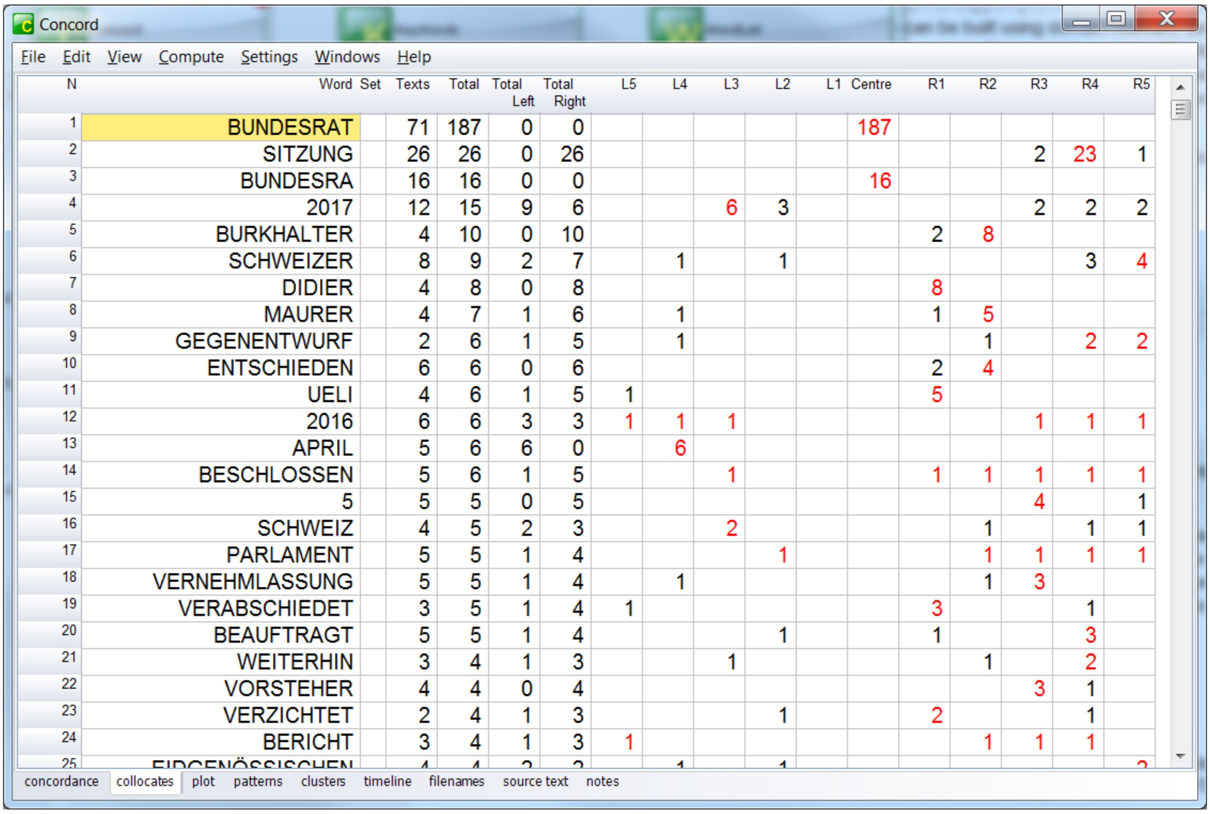
Keywords can tell us a lot about terminologisation processes: if a 'word' is particularly frequent in our corpus while having a normal distribution in reference texts, it could become a term or have a key role in the subject field we are working on. Some corpus analysis tools have a function called 'detailed consistency' that reveals if a term is specific to a particular text or text type. This tool makes it explicit which synonym should be used at a certain language register or in a specific text type. This information is particularly useful for translators and other users of terminological entries. Analysing collocates in the concordance lines helps to detect concept variation and meaning shifts as well as identifying homonyms. Concordance may also be exploited to extract contexts showing particular collocates or patterns, phraseologies as well as different kinds of definitions.

Last but not least, corpora can be imported from the internet to support processes such as terminology on demand and terminology counselling. Corpora can be built using domain-relevant 'seed' words (also multi-word expressions) as inputs via automatic queries. This procedure, called 'bootstrapping', saves time and is much more effective than manual construction of web-based corpora via time-consuming downloads. This is particularly true in the case of single-use, ad-hoc

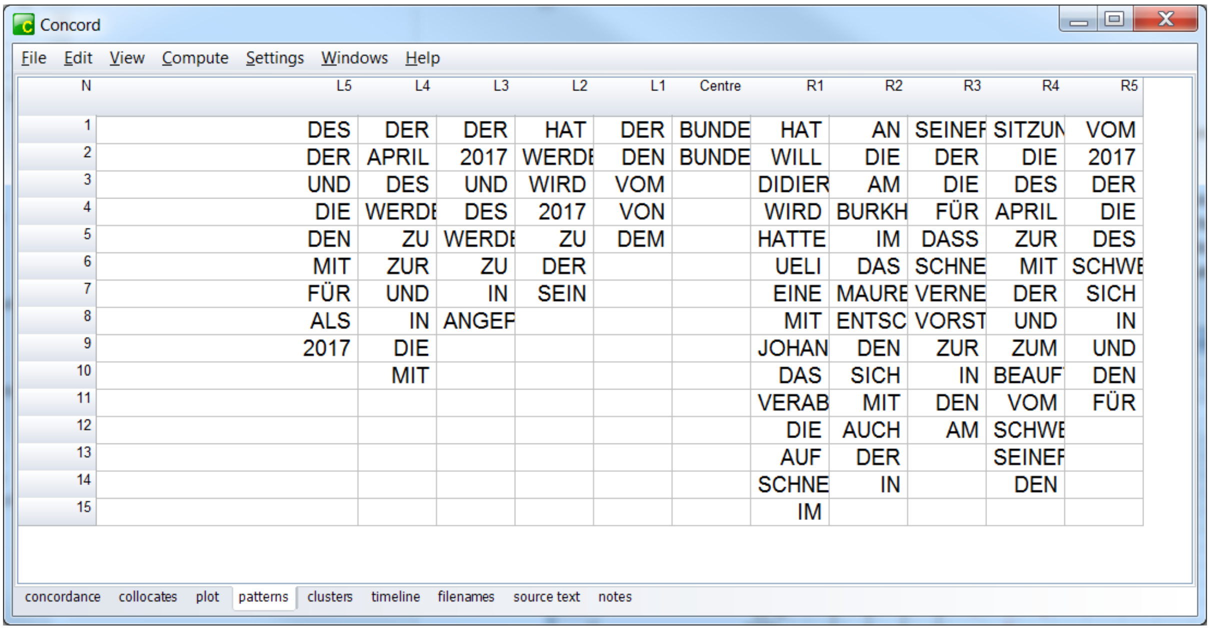
[1] In this paper, explanations of how to search corpora refer to the commercial software 'Wordsmith Tools'. Nevertheless, similar functions and procedures may work with other tools.

corpora for a specific purpose (e.g. answering terminology related questions, assessing the spread of a neologism, and so forth).

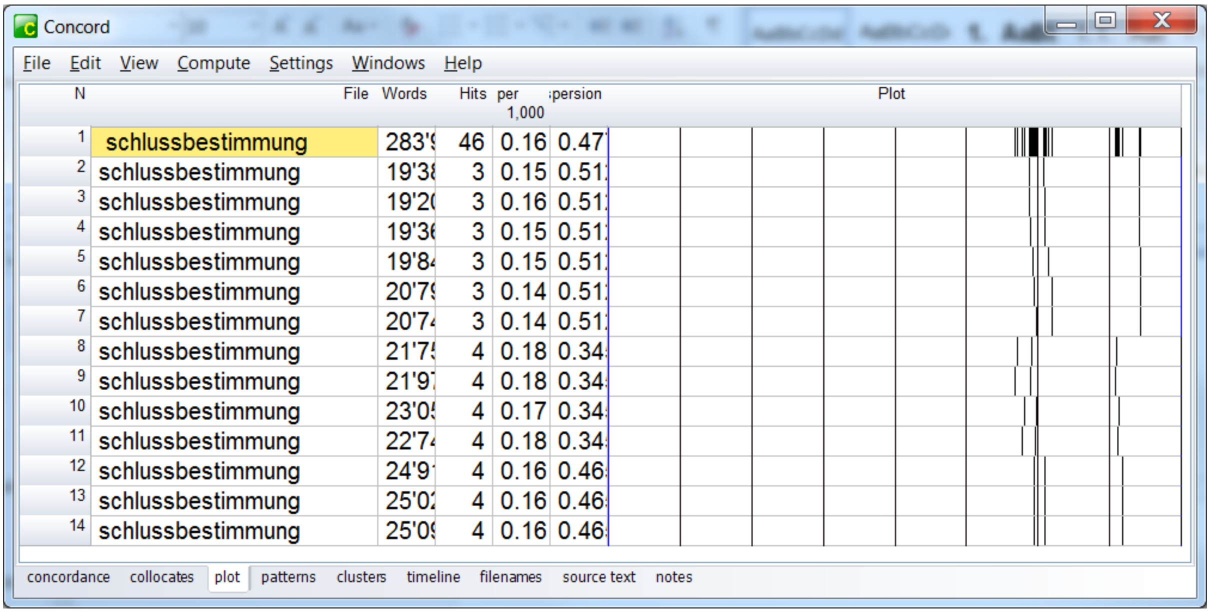
When working with parallel texts (that is, texts and their translations), some corpus features may also be analysed with a translation workbench. However, this tool only gives a raw picture of the language and does not give any explicit clue about clusters, language patterns and semantic prosody. Concordancers, on the contrary, provide terminologists with explicit statistical and linguistic information (see pictures).



N	Word	Set	Texts	Total	Total Left	Total Right	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
1	BUNDESRAT		71	187	0	0						187					
2	SITZUNG		26	26	0	26									2	23	1
3	BUNDESRA		16	16	0	0						16					
4	2017		12	15	9	6			6	3					2	2	2
5	BURKHALTER		4	10	0	10							2	8			
6	SCHWEIZER		8	9	2	7		1		1						3	4
7	DIDIER		4	8	0	8							8				
8	MAURER		4	7	1	6		1					1	5			
9	GEGENENTWURF		2	6	1	5		1						1		2	2
10	ENTSCHIEDEN		6	6	0	6							2	4			
11	UELI		4	6	1	5	1						5				
12	2016		6	6	3	3	1	1	1						1	1	1
13	APRIL		5	6	6	0		6									
14	BESCHLOSSEN		5	6	1	5			1				1	1	1	1	1
15	5		5	5	0	5									4		1
16	SCHWEIZ		4	5	2	3			2					1		1	1
17	PARLAMENT		5	5	1	4				1				1	1	1	1
18	VERNEHMLASSUNG		5	5	1	4		1						1	3		
19	VERABSCHIEDET		3	5	1	4	1						3			1	
20	BEAUFTRAGT		5	5	1	4				1			1			3	
21	WEITERHIN		3	4	1	3			1					1		2	
22	VORSTEHER		4	4	0	4									3	1	
23	VERZICHTET		2	4	1	3				1			2			1	
24	BERICHT		3	4	1	3	1							1	1	1	
25	EIDGENÖSSISCHEN		4	4	2	2		1		1							2



N	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
1	DES	DER	DER	HAT	DER	BUNDE	HAT	AN	SEINEF	SITZUN	VOM
2	DER	APRIL	2017	WERDE	DEN	BUNDE	WILL	DIE	DER	DIE	2017
3	UND	DES	UND	WIRD	VOM		DIDIER	AM	DIE	DES	DER
4	DIE	WERDE	DES	2017	VON		WIRD	BURKH	FÜR	APRIL	DIE
5	DEN	ZU	WERDE	ZU	DEM		HATTE	IM	DASS	ZUR	DES
6	MIT	ZUR	ZU	DER			UELI	DAS	SCHNE	MIT	SCHWE
7	FÜR	UND	IN	SEIN			EINE	MAURE	VERNE	DER	SICH
8	ALS	IN	ANGEF				MIT	ENTSC	VORST	UND	IN
9	2017	DIE					JOHAN	DEN	ZUR	ZUM	UND
10		MIT					DAS	SICH	IN	BEAUF	DEN
11							VERAB	MIT	DEN	VOM	FÜR
12							DIE	AUCH	AM	SCHWE	
13							AUF	DER		SEINEF	
14							SCHNE	IN		DEN	
15							IM				


Concord
File Edit View Compute Settings Windows Help
concordance collocates plot patterns clusters timeline filenames source text notes

N	File	Words	Hits	per 1,000	;persion	Plot
1	schlussbestimmung	283'9	46	0.16	0.47	
2	schlussbestimmung	19'38	3	0.15	0.51	
3	schlussbestimmung	19'20	3	0.16	0.51	
4	schlussbestimmung	19'36	3	0.15	0.51	
5	schlussbestimmung	19'84	3	0.15	0.51	
6	schlussbestimmung	20'79	3	0.14	0.51	
7	schlussbestimmung	20'74	3	0.14	0.51	
8	schlussbestimmung	21'75	4	0.18	0.34	
9	schlussbestimmung	21'97	4	0.18	0.34	
10	schlussbestimmung	23'05	4	0.17	0.34	
11	schlussbestimmung	22'74	4	0.18	0.34	
12	schlussbestimmung	24'91	4	0.16	0.46	
13	schlussbestimmung	25'02	4	0.16	0.46	
14	schlussbestimmung	25'09	4	0.16	0.46	

2. Definitions

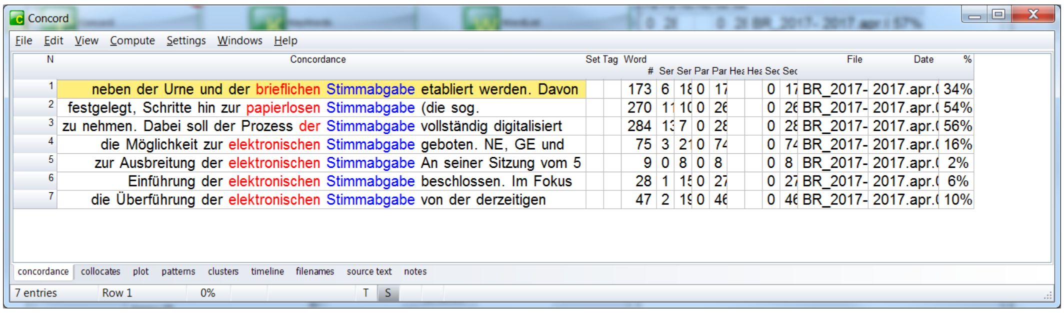
2.1. Corpus

"A collection of machine-readable authentic texts (including transcripts of spoken data) which is sampled to be representative of a particular language or language variety" (Xiao 2015: 3).

Corpora may be raw (without additional linguistic information) or annotated (tagged with additional linguistic information, typically with part-of-speech tags[2], for example *PP* for *personal pronoun*), general (reference corpora) or specialised, monolingual or multilingual, comparable or parallel, and static or dynamic (i.e. sample or monitor corpora). Static corpora do not grow over time, while dynamic corpora are constantly or regularly fed with new data. A corpus collecting Jane Austin's novels would be a static one (unless a new manuscript were found), while a corpus collecting all financial articles published in the British press would be a dynamic one, as new articles would flow in the corpus on a regular basis.

2.2. Concordance

"A set of examples of a given word or phrase, showing the context" (Scott 2017:187). Concordances help investigating collocation, colligation (see 2.4), semantic prosody and semantic preference as well as identifying collocational profiles or patterns.


Concord
File Edit View Compute Settings Windows Help
concordance collocates plot patterns clusters timeline filenames source text notes
7 entries
Row 1
0%
T
S

N	Concordance	Set	Tag	Word #	Ser	Ser	Par	Par	Hea	Hea	Sec	Sec	File	Date	%
1	neben der Urne und der brieflichen Stimmabgabe etabliert werden. Davon			173	6	18	0	17			0	17	BR_2017-	2017.apr.(	34%
2	festgelegt, Schritte hin zur papierlosen Stimmabgabe (die sog.			270	11	10	0	26			0	26	BR_2017-	2017.apr.(	54%
3	zu nehmen. Dabei soll der Prozess der Stimmabgabe vollständig digitalisiert			284	13	7	0	28			0	28	BR_2017-	2017.apr.(	56%
4	die Möglichkeit zur elektronischen Stimmabgabe geboten. NE, GE und			75	3	21	0	74			0	74	BR_2017-	2017.apr.(	16%
5	zur Ausbreitung der elektronischen Stimmabgabe An seiner Sitzung vom 5			9	0	8	0	8			0	8	BR_2017-	2017.apr.(	2%
6	Einführung der elektronischen Stimmabgabe beschlossen. Im Fokus			28	1	15	0	27			0	27	BR_2017-	2017.apr.(	6%
7	die Überführung der elektronischen Stimmabgabe von der derzeitigen			47	2	19	0	46			0	46	BR_2017-	2017.apr.(	10%

[2] See for example http://ucrel.lancs.ac.uk/annotation.html#POS for more information (last opened on 27 April 2018).

2.3. Collocation

A linear sequence of co-occurring word-forms. Documenting a term with its collocates is important especially right after term formation, as the context helps disambiguating its meaning. In many cases the collocate may complete or explain the meaning of the term. For example, knowing that *timeboxing* goes with *management* and *timeline* helps us to understand this term, even if we do not know its definition. Collocations "can be of adjective plus noun, noun plus noun, noun plus verb, etc, and […] different forms of the lemma / word family can occur in different sequences and in different spans" (Stubbs 2013: 15). They can be investigated using concordances.

According to their frequency and degree of stability, collocations may turn out to be compound terms in terminology work.

2.4. Colligation

Relationship between words co-occurring within grammatical structures (e. g. *according* and *to*) or, in a broader sense, "co-occurrence of a word with grammatical classes" (McEnery/Xiao/Tono 2006: 11).

2.5. Tokens

Running words in a text.

2.6. Types

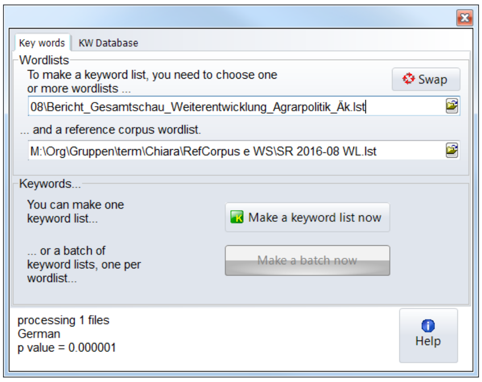
Different words in a text.

2.7. Type/Token Ratio (TTR)

A high TTR indicates that the corpus contains many different words (lexically rich); a lower TTR means that the corpus does not have many different words and could be terminologically dense. The type/token ratio is only useful if the corpus contains text having the same length; in all other cases the standardised type/token ratio (STTR) should be used. The STTR is computed every 1000 running words.

2.8. Keywords

Word forms occurring with an outstanding frequency in a text or (sub)corpus, i.e. with a higher frequency than expected and being thus 'key' for a certain text or subcorpus. Typically, keywords are obtained by comparing a specialised corpus with a reference (general) corpus.



Key words | KW Database

Wordlists

To make a keyword list, you need to choose one or more wordlists ...

08\Bericht_Gesamtschau_Weiterentwicklung_Agrarpolitik_Äk.lst

... and a reference corpus wordlist.

M:\Org\Gruppen\term\Chiara\RefCorpus e WS\SR 2016-08 WL.lst

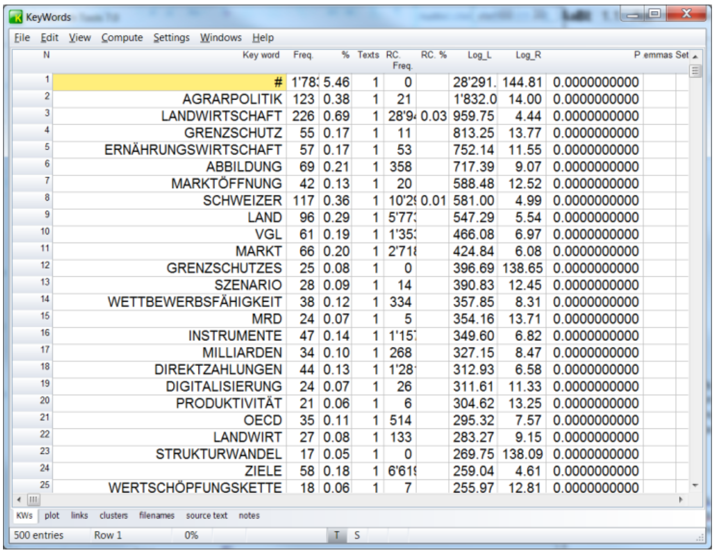
Keywords...

You can make one keyword list...

... or a batch of keyword lists, one per wordlist...

processing 1 files
German
p value = 0.000001

Swap
Make a keyword list now
Make a batch now
Help



KeyWords
File Edit View Compute Settings Windows Help
KWs plot links clusters filenames source text notes
T S

N	Key word	Freq.	%	Texts	RC. Freq.	RC. %	Log_L	Log_R	P	emmas	Set
1	#	1'78	5.46	1	0		28'291.	144.81	0.0000000000		
2	AGRARPOLITIK	123	0.38	1	21		1'832.0	14.00	0.0000000000		
3	LANDWIRTSCHAFT	226	0.69	1	28'9	0.03	959.75	4.44	0.0000000000		
4	GRENZSCHUTZ	55	0.17	1	11		813.25	13.77	0.0000000000		
5	ERNÄHRUNGSWIRTSCHAFT	57	0.17	1	53		752.14	11.55	0.0000000000		
6	ABBILDUNG	69	0.21	1	358		717.39	9.07	0.0000000000		
7	MARKTÖFFNUNG	42	0.13	1	20		588.48	12.52	0.0000000000		
8	SCHWEIZER	117	0.36	1	10'2	0.01	581.00	4.99	0.0000000000		
9	LAND	96	0.29	1	5'77		547.29	5.54	0.0000000000		
10	VGL	61	0.19	1	1'35		466.08	6.97	0.0000000000		
11	MARKT	66	0.20	1	2'71		424.84	6.08	0.0000000000		
12	GRENZSCHUTZES	25	0.08	1	0		396.69	138.65	0.0000000000		
13	SZENARIO	28	0.09	1	14		390.83	12.45	0.0000000000		
14	WETTBEWERBSFÄHIGKEIT	38	0.12	1	334		357.85	8.31	0.0000000000		
15	MRD	24	0.07	1	5		354.16	13.71	0.0000000000		
16	INSTRUMENTE	47	0.14	1	1'15		349.60	6.82	0.0000000000		
17	MILLIARDEN	34	0.10	1	268		327.15	8.47	0.0000000000		
18	DIREKTZAHLUNGEN	44	0.13	1	1'28		312.93	6.58	0.0000000000		
19	DIGITALISIERUNG	24	0.07	1	26		311.61	11.33	0.0000000000		
20	PRODUKTIVITÄT	21	0.06	1	6		304.62	13.25	0.0000000000		
21	OECD	35	0.11	1	514		295.32	7.57	0.0000000000		
22	LANDWIRT	27	0.08	1	133		283.27	9.15	0.0000000000		
23	STRUKTURWANDEL	17	0.05	1	0		269.75	138.09	0.0000000000		
24	ZIELE	58	0.18	1	6'61		259.04	4.61	0.0000000000		
25	WERTSCHÖPFUNGSKETTE	18	0.06	1	7		255.97	12.81	0.0000000000		

500 entries Row 1 0%

2.9. Seed words (or seed terms)

Words used to retrieve thematically related texts from the internet in order to build tailor-made corpora.

2.10. Cluster

In NLP tools, clusters (or lexical bundles) are words often occurring together in a fixed sequence. “They represent a tighter relationship than collocates, more like multi-word units or groups or phrases […]. These clustering relations may involve colligation […], collocation, and semantic prosody” (Scott 2017: 494). In NLP ’word clusters’ also refers to sets of paradigmatically or syntagmatically similar words (that is, synonyms or words occurring together).

2.11. Sampling

Sampling means choosing texts (= sampling units) to represent a language variety (= the ’population’). When building a sample (= the corpus), both qualitative (text types, genres, etc.) and quantitative (length, amount, etc.) aspects must be taken into account, in order to ensure that the sample/corpus – although smaller in size – has the same characteristics of the language variety being examined (population). Sampling may be random or stratified. In stratified samples the language variety to represent is divided in homogeneous groups according to the purposes for which the corpus is built (e.g. text types, authors, etc.). “Stratified samples are almost always more representative than non-stratified samples” (McEnery/Xiao/Tono 2006: 127).

Sampling is one of the first steps in systematic terminology projects. Corpus texts have to be selected and, if necessary, pre-processed if only some parts of the texts have to be used (this is the case, for instance, of texts having an introduction that may affect the statistical data such as TTR and relative frequency while not containing any interesting terminology).

2.12. Corpus balance

A corpus is balanced if it “covers a wide range of text categories which are supposed to be representative of the language or language variety under consideration. These text categories are typically sampled proportionally […]” (McEnery/Xiao/Tono 2006: 16). Corpus balance generally relies on how texts are classified and how the corpus strata are built. When classifying the texts, both linguistic and extra-linguistic features have to be taken into account.

2.13. Representativeness

"A corpus is thought to be representative of the language variety it is supposed to represent if the findings based on its contents can be generalized to the said language variety" (Leech 1991: 27). However, when working with corpora, representativeness is always a relative concept as it relates with the questions to be answered by using the corpus.

A general corpus is said to be representative if it is sampled in a balanced way. Representativeness of specialised corpora may be measured by closure (or saturation): "Closure for a particular feature of a variety of language means that it is tending towards being finite, for example, it is possible to develop a simple list of sentence type rules for the language" (McEnery/Wilson 2001: 148). However, closure may only apply to highly specialised language varieties or to narrowly defined domains (for example cardiology). In all other cases representativeness is ensured by sampling.

In most cases terminologists deal with small-sized corpora that represent highly specialised communicative situations. If corpora are very small, representativeness may be reached by saturation, but the amount of data may not be sufficient to compute relevant statistical information.

2.14. Semantic preference

We have semantic preference when a word frequently co-occurs with a set of semantically related words (paronyms). "The semantic preference is built up by observing and grouping single instances of words with similar meanings, or which in context appear to form a coherent group" (Philip 2013: 245). Collocational profiles may highlight semantic preferences and therefore help to better understand how paradigmatically related synonyms are used and the subtle differences in their intensions (in case of concepts having prototype structure).

2.15. Semantic prosody

Motivation for using the text segment in that moment (illocutionary force and discourse management) (see Stubbs 2013: 25).

3. How to build a corpus

3.1. Identifying goals and parameters

- What kind of terminology project is the corpus intended for? Thematic or ongoing collection, ad-hoc terminology, terminology on demand?
- Who are the targeted users of the terminology collection?
- Monolingual or multilingual?
- Will the terminology collection be used for descriptive or rather for prescriptive purposes?

3.2. Pre-processing

3.2.1. Text conversion (DOC, PDF, … > Unicode TXT)
3.2.2. Saving samples in an efficient folder structure (for example, text type > language…)
3.2.3. Prepare or download a stopword list for each language and load it in the corpus analysis tool. Stopword lists filter out grammatical words such as articles and prepositions that cause 'noise' while analysing the corpus.
3.2.4. Tagging (if necessary)

3.2.5. Prepare or download a lemma list for each language and load it in the corpus analysis tool. Lemma lists allow grouping different forms of a same term – such as singular and plural forms – under the same lemma in the wordlist. As a result, statistics and concordances are more accurate.

3.3. Building a corpus for a thematic collection

3.3.1. Sampling

Stratified sampling using different text types helps to identify variants of the preferred terms and sets of synonyms and to understand if they are used in all text types or only in some of them.

3.3.2. Size

Given domain restriction and tendency to saturation, a small corpus (about 100.000 tokens) may work better than a big one for corpus-based terminography.

3.4. Building a corpus for an ongoing terminology collection

3.4.1. Sampling

Ongoing terminology collections require dynamic, ever growing corpora. Depending on the terminology project, it may be useful to work with a reference corpus and various subcorpora, which may be monitoring corpora (if terminology changes or shifts need to be detected) or corpora pertaining to a subdomain. The reference corpus may need to be refreshed (new subcorpora may flow into the reference corpus after a while or when the next project batch is to be processed). For instance, using legislation as a reference corpus requires adding new normative texts to the corpus as soon as they are issued.

3.4.2. Size

In this case, the size of the subcorpora being added to keep the corpus up-to-date may not be decided by the corpus compiler. Thus, it is important to adjust the program settings according to corpus size in order to ensure efficiency of the analysis (the association or keyness measure may have to be selected according to corpus size[3] as the latter affects the identification of collocations and keywords).

3.5. Building a corpus for ad-hoc purposes

When ad-hoc terminology tasks have to be performed (consultancy, answering specific questions by third parties, carrying out researches to solve specific, text-related terminology problems, etc.), efficiency in terms of time, costs and results is one of the most important issues. Bootstrapping a corpus from the internet may help to meet these requirements. This method saves time and allows to build tailor-made corpora from scratch in a few minutes. Problematic terms and related terms can be used as seed words; typically, bootstrapping tools allow to adjust settings so as to narrow the retrieving task (e.g. by limiting the search to given domain extensions or excluding extensions) and offer various corpus analysis possibilities.

Depending on the task to be performed, superordinate, subordinate or coordinate terms may be chosen as seed words.

4. How to exploit corpora in terminography…

[3] MI^2 usually performs best, see also 4.3. On this topic see for example http://ucrel.lancs.ac.uk/llwizard.html (last opened on 30th April 2018).

4.1. …to select candidate terms

Simple terms: Wordlist

Complex terms (multiword expressions): Wordlist > Index[4] > Compute clusters

When working with a reference corpus, at this stage any comparison should be done between wordlists and not with keyword lists: this ensures that all the words in both lists are compared. The procedure yields all words which appear significantly more often in one than the other corpus, including those which appear more than a minimum number of times in one corpus even if they do not appear at all in the other one.

Wordlist > Compare two wordlists

4.2. …to identify terms

The identification of terms to be included in a thematic collection is based on the candidate terms selected at a first stage. This second selection may involve an assessment of the terms' keyness (both positive and negative, > Keywords and p value). In Wordsmith Tools, the keyness of complex terms must be assessed by comparing wordlists based on word clusters.

Another useful tool to identify terms among candidates is consistency analysis. This allows to understand whether the term recurs consistently in many texts. Scott (2015) gives a useful example for terminology: "For example, the word *consolidate* was found to occur in many of a set of business Annual Reports. It did not occur very often in each of them, but did occur much more consistently in the business reports than in a mixed set of texts." Given the fact that terms may have low frequency (which is statistically less significant) and yet a high relevance for a given domain or text type, consistency analysis helps assess the term relevance. Consistency analysis may be plain or detailed.

The identification of terms also includes the structuring of candidate terms in a concept system as well as disambiguation of synonyms and variants. To detect synonyms and variants: concordance > "which is/i.e./also called…".

To disambiguate between variants and identify the preferred term, ConcGrams is a useful tool: it allows to identify non consecutive linkages thus recognising morphosyntactic variants of a candidate term.

4.3. …to extract complex terms

To get an overview: Wordlist > Index > Compute clusters

To detect subordinate terms having the form of noun phrases in generic concept systems:

Concordance > 'term' > collocates > language-specific sorting (for English and German L1, for Italian and French R1). According to Bartsch/Evert (2013: 31-33), the best performing association measure in this case would be MI^2. Log-likelihood (LL) overestimates high frequency words while neglecting others, while the Dice coefficient is less sensitive to frequent words and highlights topical and informative words. Wordlist > Index > compute > relationships gives an overview of how strong the association between a term and another word is.

[4] In Wordsmith Tools, an index is a wordlist that records "the positions of all the words in your text file, so that you can subsequently see which word came in which part of each text" (Scott 2017: 304). Indexes are used to compute clusters and to investigate how word types relate to each other (mutual information score).

To detect subordinate terms of a given term:

Concordance > Clusters (tab) (yields multi-word expressions containing the term, thus also MWE having the structure modifier-head)

Concordance > Compute (menu item) > Clusters

Concordance > Patterns

4.4. …to investigate collocations

For terminology work, collocations computed with a concordancer can serve various purposes:

1) **to understand the degree of specialisation of a term**. Collocations help to understand in which register and in which kind of discourse a term is used as they may vary from "various general and preferred use, 'loosen/remove the screw'" (Champe in Wright/Budin 2001:511);
2) **to identify phraseology**. Collocational patterns allow detecting both consecutive and non-consecutive linkages. This tool therefore helps to investigate phraseologies subject to morphosyntactic variation in their inner structures. This feature is especially useful if using a tagged corpus, as POS-tags may be used as search strings, too. For the same reason, collocational patterns may be used to investigate terms prone to expansion and reduction phenomena. Concgrams may be used for the same purpose;
3) **to check terminological cohesion of a text**. Identifying an inventory of collocations and collocates for a given term helps to assess the terminological cohesion of a text. While CAT-tools usually allow for checking consecutive multiword expressions if they have been saved for the terminology check, a concordancer makes it possible to check collocates even if they are inflected or separated from the search term by a wide span, provided a sufficiently broad collocational horizon has been set.

4.5. …to disambiguate synonyms

To investigate the degree of synonymy between two terms, collocations and collocates may help, as they give information about register, discourse, and characteristics of a given term. In particular, comparing collocate verbs helps to better understand the intension of a concept. For instance, in a corpus of family law the collocational profiles of *amministratore* (*administrator*) and *curatore* (*deputy*) only share the collocate *nomina* (*appointment*); however, this collocate occurs in different positions (more frequently L2 in relationship with *curatore* and L3 in relationship with *amministratore*). Among the collocates of *amministratore* there are no active verbs, while a *curatore* performs different actions: *deve*, *adempie*, *acquisisce*, *acconsente*, *rimette*, *rappresenta*, *redige*, *informa* and so on; indeed, the two terms are not synonyms at all, as they perform different functions.

The dispersion plot is another feature that helps to disambiguate synonymic relationships between terms. In Wordsmith Tools, the dispersion plot "shows where the search word occurs in the file which the current entry belongs to. That way you can see where mention is made most of your search word in each file" (Scott 2015: 203). For instance, in a corpus of government financial statements, the plots of the terms *receipts* and *revenues* look quite different. Indeed, they occur at different positions in the text, as *receipts* and *expenditure* define concepts relating to the financing statement, whereas *revenue* and *expenses* are used in the statement of financial performance. In this context, these terms are not synonyms.

4.6. …to explore contexts, definitions and variants

Corpora are an excellent empiric source for contexts as examples of language use. Concordances, collocates and patterns help to understand how terms are used in texts and select the most relevant contexts to add to the entry.

Concordances may also be used to look for definitions and variants by entering some given expressions as search strings:

- '*that is*' may yield paraphrases, which can be used as definitions or notes in the terminological entry
- '*is/are made of*' yields partitive definitions
- '*term + is*' may yield paraphrases or explanations, which can be used as definitions or notes in the terminological entry
- '*Under + term + we understand*' and similar expressions may yield terminological definitions based on hierarchical relationships, thus containing superordinate and subordinate terms.
- '*also called/also known as/…*' and similar expressions may yield variants or synonyms.
- …and so forth.

4.7. …to explore linguistic patterns

In addition to patterns of lexical collocations, corpora also deliver information about grammatical collocations, that is, colligations. This is useful for comparing term structure, government and valency across different languages. Any relevant difference can be described in the note of the terminological entry.

Colligations can be investigated by calculating relationships between collocates:

Concordance > Search term > collocates > Compute > relationships.

Collocates can then be sorted according to their position in order to detect which word occurs most frequently with the search term in a given position. For instance, the Italian verb *acconsentire* strongly collocates with the preposition *a* in R1 position. A tagged and lemmatised corpus allows optimal use of these features.

4.8. …to identify neologisms

For this purpose we may define a neologism as "an observable instance of change in the data which is not due to noise or accidental variation, but rather reflects what could reasonably be thought to indicate diachronic change in the language of which the corpus is a sample" (Bürki 2013: 42).

Consequently, five types of change may be identified (according to Bürki 2013: 42):

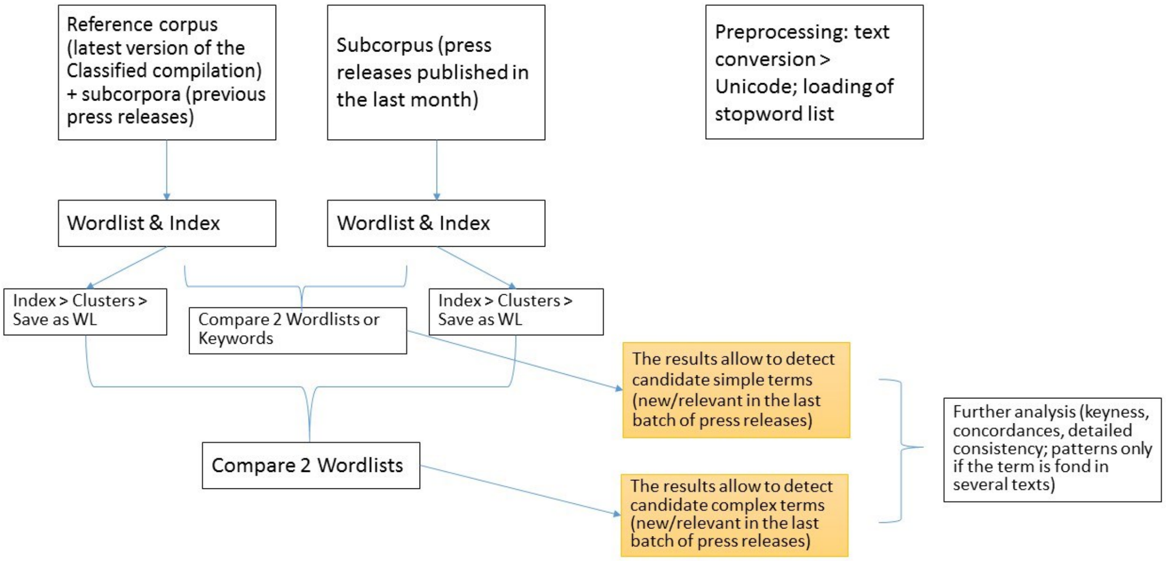
- Appearance of new types
- Disappearance of existing types
- Semantic shifts (terminologisation or de-terminologisation of existing types)
- Change in form (stable semantic) > new variants
- Notable in- and/or decreases in frequency (may indicate a changes in the extra-linguistic reality).

All these neologisms may be detected analysing statistical information such as relative frequency and keyness as well as studying collocational profiles, provided we have an up-to-date monitor corpus to work with.

5. Case study

5.1. Using corpora to feed non-systematic, on-going terminology collections

Terminology of the press releases of the Swiss Federal Administration.


Reference corpus (latest version of the Classified compilation) + subcorpora (previous press releases)
Subcorpus (press releases published in the last month)
Preprocessing: text conversion > Unicode; loading of stopword list
Wordlist & Index
Wordlist & Index
Index > Clusters > Save as WL
Compare 2 Wordlists or Keywords
Index > Clusters > Save as WL
The results allow to detect candidate simple terms (new/relevant in the last batch of press releases)
Compare 2 Wordlists
The results allow to detect candidate complex terms (new/relevant in the last batch of press releases)
Further analysis (keyness, concordances, detailed consistency; patterns only if the term is fond in several texts)

The following process is performed with *Wordsmith Tools*. Ideally, after having selected terms among candidates, the terms should be used as seeds for bootstrapping specific corpora from relevant websites (in our case, admin.ch and other institutional pages) in order to double-check the terms extracted from our corpus, to add variants and synonyms, and to retrieve definitions and contexts if they are not provided in the press release or if better sources are found. In absence of the necessary tools, these steps are performed via Google.

6. Some Tools

6.1. AntConc

AntConc is a "freeware corpus analysis toolkit for concordancing and text analysis" (see website http://www.laurenceanthony.net/software.html) developed by Dr. Laurence Anthony. It does not need to be installed and runs on Windows, Macintosh and Linux. In addition to AntConc, other freeware resources for natural language processing can be downloaded from the same page.

6.2. WordSmith Tools

WordSmith Tools is a commercial software (available for Windows); it is "an integrated suite of programs for looking at how words behave in texts" (Scott 2015: 2). Is includes three main tools (WordList, Concord, KeyWords) and many other features such as Textconverter, ConcGrams, and so forth. "The tools have been used by Oxford University Press for their own lexicographic work in preparing dictionaries" (Scott 2015: 2). Wordsmith Tools needs local installation; corpora should also be saved locally (or on a server, although this may lead to slower performance).
http://www.lexically.net

6.3. Sketch Engine

Web-based, commercial software for bootstrapping corpora from the internet. It also allows to search the corpus for concordances, collocates and so forth, as well as to extract wordlists. Furthermore, SketchEngine gives access to larger general language corpora (including the British National Corpus, the Brown Corpus, the EUROPARL parallel corpora and the TenTen Web Corpora), which can be used as reference corpora when computing keywords in LSP corpora.

The specific features of the Sketch Engine are “word sketches, one-page, automatic, corpus-derived summaries of a word's grammatical and collocational behaviour” (see website http://www.sketchengine.co.uk/).

6.4. BootCat

Freeware for bootstrapping corpora from the internet. Different versions are available for download (compliance with the main operative systems). BootCat is available as a toolkit (command-line scripts for advanced use) or as front-end version, “which is a graphical interface for the BootCaT toolkit [. I]t's basically just a wizard that guides you through the process of creating a simple web corpus. The front-end does not yet support all the features available in the command-line scripts, advanced users comfortable with text UIs should consider using the Perl scripts instead of the front-end”, see website http://bootcat.sslmit.unibo.it/).

7. Open source resources

- Tree Tagger: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ (POS Tagger for German, English, French, Italian, Dutch, Spanish, Bulgarian, Russian, Portuguese, Galician, Chinese, Swahili, Slovak, Latin, Estonian, Polish and old French; works as a chunker for English, German, and French).

- Stopword lists and stemmers: http://members.unine.ch/jacques.savoy/clef/index.html

- Profiles of co-occurrence (German only): http://corpora.ids-mannheim.de/ccdb/

8. References

- R. Xiao. 2015. *Corpus design and types of corpora*. See http://www.lancaster.ac.uk/fass/projects/corpus/ZJU/xpresentations/session%202.ppt (last accessed 2017.07.27).
- M. Scott. 2017. WordSmith Tools Manual. Version 7.0. See http://lexically.net//downloads/version7/help_manual.zip (last accessed 2017.07.27).
- 7th International Corpus Linguistics Conference. Abstract Book. Lancaster 23-26 July 2013. (contains quoted abstracts by Stubbs, Bürki, Philip, Cvrček / Fidler, Bartsch/Evert).
- Firth, J. R. [1951]. 1957. ’Modes of Meaning’. In: Papers in Linguistics 1934-1951. Oxford: Oxford University press, 190-215.
- T. McEnery, R. Xiao and Y. Tono. 2006. *Corpus-based Language Studies: An Advanced Resource Book*. London and New York: Routledge.
- M. Stubbs. 2013. ’Sequence and order. The neo-firthian tradition of corpus semantics'. In H. Hasselgård, J. Ebeling & S. Oksefjell Ebeling (eds.) *Corpus Perspectives on Patterns of Lexis*. Amsterdam: Benjamins, 13-33.
- G. Philip. 2013. ’ A defence of semantic preference’. In 7th International Corpus Linguistics Conference. Abstract Book, pp. 244-246.