Semi-automatic Evaluation of Terminological Web-crawled Corpora

Lotte Weilgaard Christensen

University of Southern Denmark, Kolding, Denmark
lotte@sdu.dk

Abstract. This paper presents a method for evaluating the suitability of web-crawled corpora for terminological analyses. Since the contents of web-crawled corpora are unknown, the need arises for testing whether such corpora comprise a sufficient quantity of terminological information, the focus being on linguistic and conceptual coverage. The analyses are based on the corpus management system Sketch Engine, combined with knowledge patterns based on the valency of Danish verbs. The results originate from corpora established for exam purposes and used by students. So far Sketch Engine has been used primarily by lexicographers. However, the paper demonstrates that terminologists may use the program for far more than term extraction.

Keywords: corpus evaluation; web-crawled corpora; corpus tools; Sketch Engine; knowledge patterns; terminology extraction; knowledge extraction

1 Introduction

This paper aims to discuss methods for evaluating to what extent a web-crawled corpus compiled from the Internet comprises information of relevance for terminology work. The challenge to be faced by users of such a corpus is the fact that they do not know its contents, and that consequently, there is a risk of performing terminology work on an unknown basis. In the article ‘Getting to know your corpus’, aimed at lexicographical investigations, Adam Kilgarriff [6] raises some questions of great relevance in this connection: “But can we trust a crawled corpus?”, and “How do we know what is in it, or if it does a good job of representing the language?”

However, knowing if the amount of knowledge represented by the linguistic data of a given corpus is sufficient for our investigation is necessary for terminological investigations. Even if a domain corpus comprises a large number of terms, it will not necessarily include sufficient terminological information to identify the semantic relations required to establish conceptual systems, nor will it necessarily comprise linguistic data suitable as input for definitions.

The web-crawled corpora will be tested using the corpus management system Sketch Engine, widely used by lexicographers [5]. To my knowledge, in connection with terminological investigations, Sketch Engine has primarily been used for extracting term candidates [5], whereas little attention has been given to the extraction of other terminological information by means of the system. The article will demonstrate that Sketch Engine is far more capable of supporting terminology work than what has been described until now. Moreover, the Sketch Engine team is in the process of developing methods for automatic extraction of hierarchical relations as well as definitions [1]. Thus, the aim of the article is to discuss how web-crawled corpora compiled by Sketch Engine may be tested, and to demonstrate how Sketch Engine may also support the retrieval of terminological information at the conceptual level. In the semi-automatic evaluation methodology, a subset of Danish knowledge patterns have been implemented which are suitable for retrieval of knowledge-rich contexts (KRC). Knowledge-rich context has been defined by Meyer [7] as "a context indicating at least one item of domain knowledge that could be useful for conceptual analysis".

The retrieval of information from corpora and the evaluation of corpora for terminological purposes are to some extent two sides of the same coin. In this paper, focus will be on monolingual information retrieval. Besides, the approaches needed for different languages, depending on the corpora and the corpus analysis functions available will be compared. In that connection, the importance of finding simple methods easily applied by all user groups must be emphasized.

The rest of the paper is organized as follows: in Section 2, I describe the background, in Section 3, the criteria to be used for corpus design are discussed. Section 4 comprises a general description of Sketch Engine. Sections 5 and 6 deal with issues of linguistic and conceptual coverage, including Sketch Engine functions and knowledge patterns. Section 7 ends with some concluding remarks.

2 Background

This investigation focuses on specialized corpora used for exam assignments in terminology courses in which students are expected to demonstrate their mastering of the methodology of terminology. In the typical assignment, students will be asked to construct a conceptual system comprising 10 to 20 concepts and to write definitions of some of the concepts in question.

Compiling web-crawled corpora for the above purposes revealed that although they comprised a large number of term candidates, even corpora consisting of more than 50,000 tokens might not necessarily include sufficient amounts of elements of knowledge to enable students to carry out the terminology tasks required. And when students have been given a corpus for terminological investigations for the purposes of an exam, they naturally expect the corpus to be suitable for the retrieval of conceptual information and not only for the identification of terms.

3 Criteria of Corpus Design

Obviously, the work load involved in building a web-crawled corpus is smaller than the one needed to compile a well-designed corpus. However, if the corpus texts used for a terminology project turn out to be of inferior quality, the end result of the project will likewise be of inferior quality and turn out a very expensive one. Thus, since the contents of web-crawled corpora are not known, methods of evaluating such corpora must be found.

Bowker and Pearson [3] define corpus as 'a large collection of authentic texts that have been gathered in electronic form according to a specific set of criteria'. Indeed, it is generally agreed in terminology literature that well-designed corpora should be compiled according to specific criteria determined by the goals of the task in hand, criteria such as text type and function, reliability, level of expertise, and domain coverage, including both linguistic and conceptual coverage.

In what follows, a method of semi-automatic validation of corpora, focusing on the criteria of linguistic and conceptual coverage will be presented. Those criteria have been chosen because a domain-specific corpus must find itself at a level of linguistic and conceptual coverage that will suffice for the purpose defined; if not, the students will not be able to use it to demonstrate their ability to apply terminological methodology. Here the term linguistic coverage refers to a sufficient number of terms in a corpus. Likewise, the term conceptual coverage refers to a sufficient amount of knowledge-rich contexts for the purpose of a given exam assignment. As indicated earlier, the testing is based on Sketch Engine combined with Danish knowledge patterns.

4 The Corpus Management System Sketch Engine

Sketch Engine is an advanced commercial corpus management system. It primarily distinguishes itself from other corpus tools by comprising an integrated piece of software called WebBootCat, compiling texts into a corpus by crawling the web. In order to create a corpus from the Web, the user is asked to specify 3 to 20 seed words, i.e. key words or multi-word expressions, from the subject domain to be investigated [5]. In a sense, at this stage the seed words are the criteria defining your corpus.

In addition to basic functionalities known from other corpus analysis tools, the corpus analysis tool of Sketch Engine offers additional advanced functionalities, some of which will be described below.

Moreover, Sketch Engine includes large LGP corpora in sixty languages [6], to be applied as reference corpora when extracting lists of term candidates. For English, to name an example, the British National Corpus is available [5]. For Danish, large LGP corpora have also been added in recent years.

In order to apply the advanced functionalities of Sketch Engine, so-called 'high-level resources' must be integrated, including: a tokeniser, a lemmatizer, a part-of-speech tagger, and a parser or 'sketch grammar' [5]. A sketch grammar identifies possible relations of words to a keyword [8].

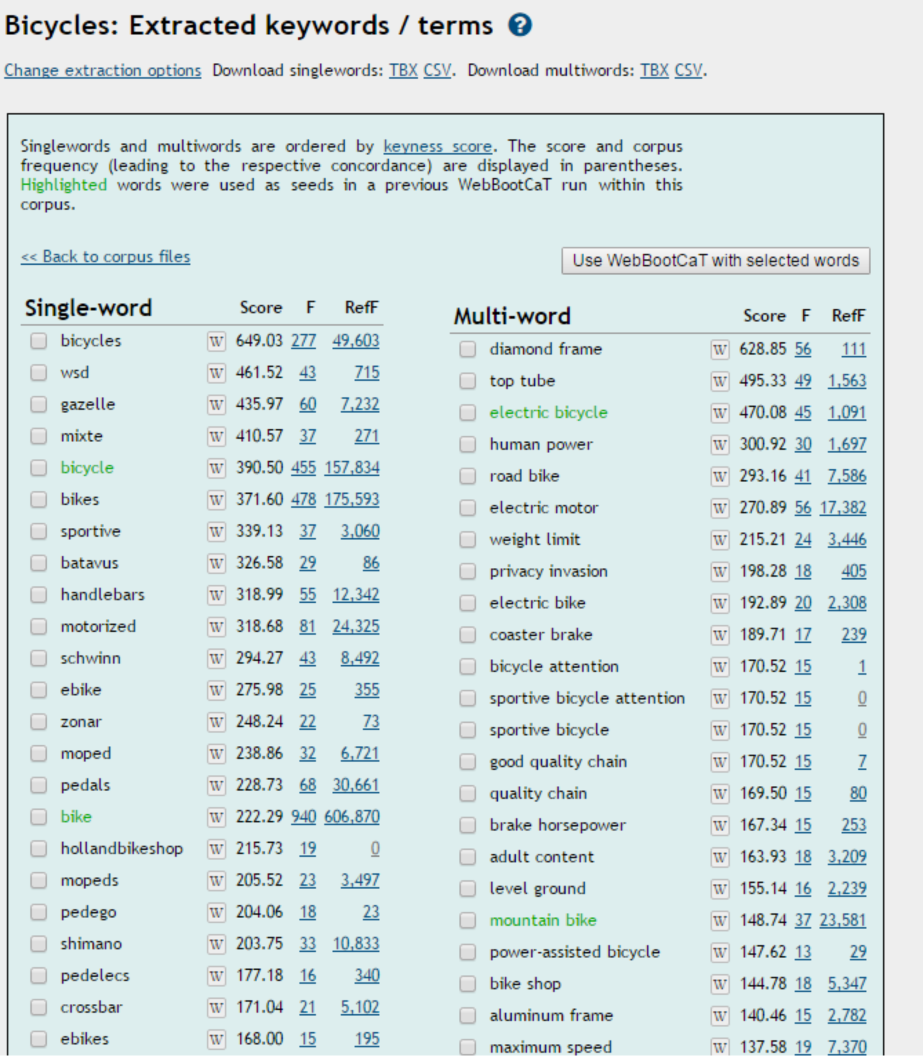
Sketch Engine does not support all languages with high-level resources. This means that depending on the language used, there will be substantial differences as to what analyses can be carried out using Sketch Engine, as well as to the ways in which it will support the terminological investigations. Since Danish is one of the languages for which high-level resources are not available, at least not for the untagged domain specific user corpora, users must rely on functions based on statistical calculations and find pragmatic approaches to information retrieval as well as to testing the usability of their corpora.

5 Linguistic Coverage

Below, it will be illustrated how the linguistic coverage of a corpus can be evaluated using Sketch Engine. The examples on *bicycles* originate from a corpus compiled for an exam assignment on this topic.

5.1 Term Extraction

The first step when testing the linguistic coverage of a corpus is to apply the term extractor function offered by Sketch Engine. This function compares the domain specific corpus to a reference corpus. The term extractor generates a file consisting of two columns called 'Single-word' (in earlier versions 'keywords') and 'Multi-word' (in earlier versions 'terms'), respectively. From a terminological perspective, the new designations are more motivated since both columns represent term candidates. The columns are shown in Fig. 1 below, retrieved from an English corpus on *bicycles*, compiled for this purpose, totaling 73,961 tokens. From the frequency information in the columns, a concordance list can be accessed directly.



Bicycles: Extracted keywords / terms

Change extraction options Download singlewords: TBX CSV. Download multiwords: TBX CSV.

Singlewords and multiwords are ordered by keyness score. The score and corpus frequency (leading to the respective concordance) are displayed in parentheses. Highlighted words were used as seeds in a previous WebBootCaT run within this corpus.

<< Back to corpus files

	Single-word		Score	F	RefF
□	bicycles	W	649.03	277	49,603
□	wsd	W	461.52	43	715
□	gazelle	W	435.97	60	7,232
□	mixte	W	410.57	37	271
□	bicycle	W	390.50	455	157,834
□	bikes	W	371.60	478	175,593
□	sportive	W	339.13	37	3,060
□	batavus	W	326.58	29	86
□	handlebars	W	318.99	55	12,342
□	motorized	W	318.68	81	24,325
□	schwinn	W	294.27	43	8,492
□	ebike	W	275.98	25	355
□	zonar	W	248.24	22	73
□	moped	W	238.86	32	6,721
□	pedals	W	228.73	68	30,661
□	bike	W	222.29	940	606,870
□	hollandbikeshop	W	215.73	19	0
□	mopeds	W	205.52	23	3,497
□	pedego	W	204.06	18	23
□	shimano	W	203.75	33	10,833
□	pedelecs	W	177.18	16	340
□	crossbar	W	171.04	21	5,102
□	ebikes	W	168.00	15	195

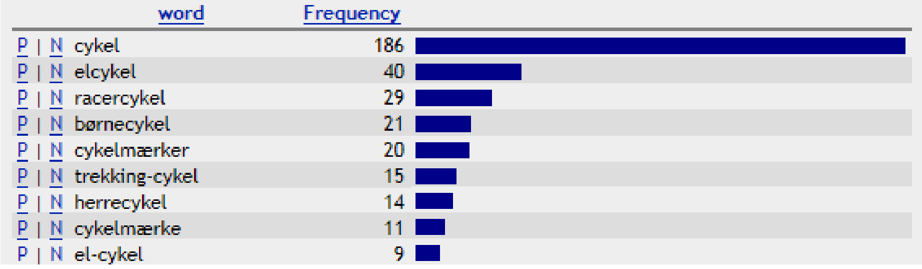
	Multi-word		Score	F	RefF
□	diamond frame	W	628.85	56	111
□	top tube	W	495.33	49	1,563
□	electric bicycle	W	470.08	45	1,091
□	human power	W	300.92	30	1,697
□	road bike	W	293.16	41	7,586
□	electric motor	W	270.89	56	17,382
□	weight limit	W	215.21	24	3,446
□	privacy invasion	W	198.28	18	405
□	electric bike	W	192.89	20	2,308
□	coaster brake	W	189.71	17	239
□	bicycle attention	W	170.52	15	1
□	sportive bicycle attention	W	170.52	15	0
□	sportive bicycle	W	170.52	15	0
□	good quality chain	W	170.52	15	7
□	quality chain	W	169.50	15	80
□	brake horsepower	W	167.34	15	253
□	adult content	W	163.93	18	3,209
□	level ground	W	155.14	16	2,239
□	mountain bike	W	148.74	37	23,581
□	power-assisted bicycle	W	147.62	13	29
□	bike shop	W	144.78	18	5,347
□	aluminum frame	W	140.46	15	2,782
□	maximum speed	W	137.58	19	7,370

Fig. 1. Term extraction in Sketch Engine (not complete)

For languages not supported by the high-level resources, the term extractor function only generates a list of single-words, i.e. only single-word term candidates may be found. This means that Danish users must apply a more pragmatic approach in order to extract multi-word terms, using the concordance function. This approach has already been described in connection with other corpus analysis tools.

However, for languages in which many concepts are represented by composite terms, the next step in evaluating the degree of linguistic coverage is to enter a generic term, e.g. in our case *cykel* as a common head, which is the Danish word for *bicy-*

cle, and to search on this term as a truncated character string in order to identify potential types of the generic term which may indicate subordinate concepts. This search result may be re-sorted alphabetically using the search node so that all instances of the same composite terms are grouped together. On the basis of the concordance list sorted by node form, it is possible to generate a frequency list of the node forms in question. In this way, the search result from the concordance list can be narrowed down, making it easier to use than a multi-page concordance list, as illustrated in Fig. 2:



	word	Frequency
P | N	cykel	186
P | N	elcykel	40
P | N	racercykel	29
P | N	børnecykel	21
P | N	cykelmærker	20
P | N	trekking-cykel	15
P | N	herrecykel	14
P | N	cykelmærke	11
P | N	el-cykel	9

Fig. 2. Extract of frequency list of composite term candidates including *cykel* as generic term

The next step will be to search for the generic term without truncation, and to sort the search result using the word strings immediately to the left hand-side and immediately to the right-hand side of the generic term, respectively, in order to identify recurrent patterns that might represent multi-word terms, such as *elektrisk cykel (electric bicycle).*

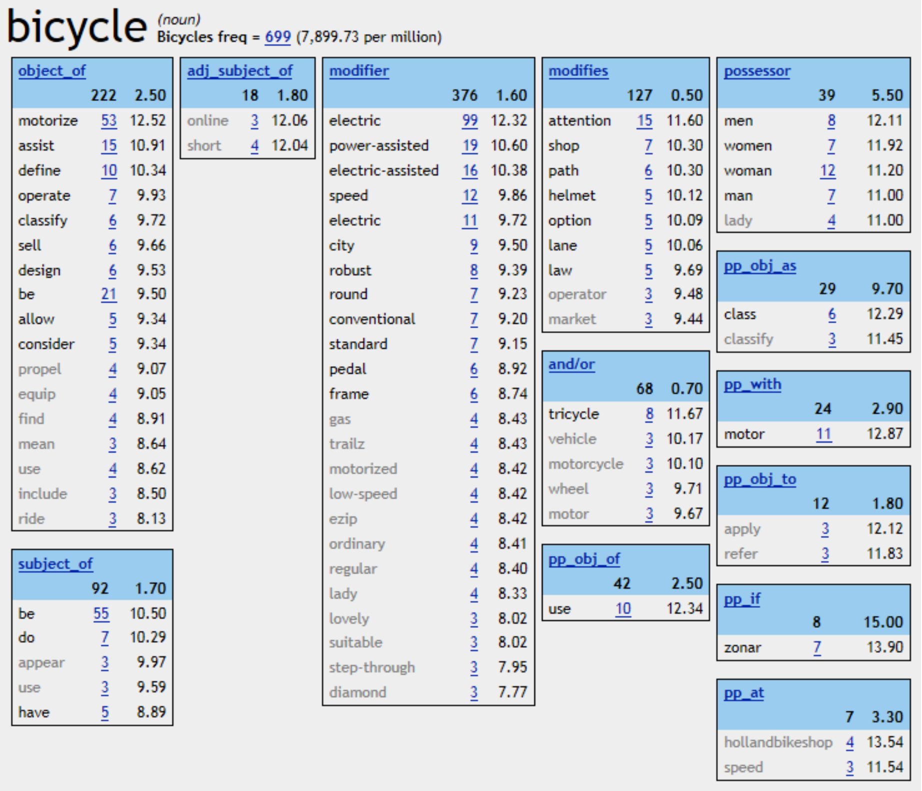
5.2 Word Sketch Used for Term Extraction

Sketch Engine has its name from the word sketch which is a core function consisting of a one-page summary of a specific word's grammatical and collocational behaviour [5]. In other words, the word sketch is a list containing different recurrent patterns of the word searched for, according to the grammatical function of the word. The word sketch function requires a sketch grammar, cf. section 4. Fig. 3 shows a word sketch for the noun *bicycle*, retrieved from the English corpus on *bicycles.*

The columns labeled ‘modifier’ and ‘modifies’ offer information on term candidates. In the first case, *bicycle* is the head of potential multi-word terms with modifiers such as *electric, electric-assisted*, *power-assisted*, or *city*. In the second case *bicycle* is the modifier in nouns such as *bicycle shop, bicycle helmet*.

Compared to the manual work that must be carried out by the terminologist analyzing concordances, the word sketch provides him or her with a quick and easy overview of the recurrent patterns in which multi-word term candidates may occur.

Any word occurring as a frequent word together with the word searched for in word sketch will be provided with a frequency number. Via this number the relevant concordance list can be accessed directly.



bicycle (noun)
Bicycles freq = 699 (7,899.73 per million)

object_of	222	2.50
motorize	53	12.52
assist	15	10.91
define	10	10.34
operate	7	9.93
classify	6	9.72
sell	6	9.66
design	6	9.53
be	21	9.50
allow	5	9.34
consider	5	9.34
propel	4	9.07
equip	4	9.05
find	4	8.91
mean	3	8.64
use	4	8.62
include	3	8.50
ride	3	8.13

subject_of	92	1.70
be	55	10.50
do	7	10.29
appear	3	9.97
use	3	9.59
have	5	8.89

adj_subject_of	18	1.80
online	3	12.06
short	4	12.04

modifier	376	1.60
electric	99	12.32
power-assisted	19	10.60
electric-assisted	16	10.38
speed	12	9.86
electric	11	9.72
city	9	9.50
robust	8	9.39
round	7	9.23
conventional	7	9.20
standard	7	9.15
pedal	6	8.92
frame	6	8.74
gas	4	8.43
trailz	4	8.43
motorized	4	8.42
low-speed	4	8.42
ezip	4	8.42
ordinary	4	8.41
regular	4	8.40
lady	4	8.33
lovely	3	8.02
suitable	3	8.02
step-through	3	7.95
diamond	3	7.77

modifies	127	0.50
attention	15	11.60
shop	7	10.30
path	6	10.30
helmet	5	10.12
option	5	10.09
lane	5	10.06
law	5	9.69
operator	3	9.48
market	3	9.44

and/or	68	0.70
tricycle	8	11.67
vehicle	3	10.17
motorcycle	3	10.10
wheel	3	9.71
motor	3	9.67

pp_obj_of	42	2.50
use	10	12.34

possessor	39	5.50
men	8	12.11
women	7	11.92
woman	12	11.20
man	7	11.00
lady	4	11.00

pp_obj_as	29	9.70
class	6	12.29
classify	3	11.45

pp_with	24	2.90
motor	11	12.87

pp_obj_to	12	1.80
apply	3	12.12
refer	3	11.83

pp_if	8	15.00
zonar	7	13.90

pp_at	7	3.30
hollandbikeshop	4	13.54
speed	3	11.54

Fig. 3. Word sketch for *bicycle*

For Danish, the word sketch function is not available for crawled user corpora. At present, for Danish or other languages without high-level resources, it is necessary to retrieve the information comprised by the word sketch by analyzing the concordance lists manually.

In his article ‘Getting to know your corpus’, Kilgarriff [6] argues that keyword lists, combined with a Sketch Engine function comparing two corpora, based on a model called simple math, is an essential support for the user wanting to gain an overview of the contents of a corpus, since in Kilgarriff’s words, a keyword list “takes frequency lists as summaries of the two corpora, and shows us the most contrasting items” [6]. This is true as far as linguistic coverage is concerned. However, this will not suffice for terminological investigations.

6 Conceptual Coverage

For terminological investigations, we obviously need a method to secure sufficient conceptual coverage as well. The next natural step will be to identify possible relations among concepts in order to be able to work out preliminary drafts of concept systems. The searches mentioned above for generic terms constituting the shared heads of composite terms or of multi-word terms will give you an impression, not just of the degree of linguistic coverage, but frequently also of potential terms representing subordinate concepts entering into type relations with the generic term (concept). However, for students to be able to carry out thorough terminological investigations, it must be secured that the corpus contains explicit knowledge-rich contexts.

6.1 Knowledge Patterns for Danish

Previously, I have analyzed domain specific corpora with the object of identifying knowledge patterns for Danish, my main focus being on recurrent patterns of verbs and their surroundings. My approach was originally based on a valency theory called the Pronominal Approach [4], building mainly on syntactic criteria. Many Danish verbs are formed analytically by means of e.g. prepositional objects and particles, as described in Weilgaard Christensen [9,10]. Thus, the approach in question enables identification of search patterns consisting of a verb together with a specific preposition. This character string approach makes it possible to eliminate terminologically irrelevant patterns (noise), and thus to narrow down the search result. As a natural consequence, an important insight achieved is that optional arguments which occur with prepositions become mandatory when they are used for the retrieval of terminological data [9,10].

In fact, for some verbs it is possible to predict with a considerable degree of certainty which terminological information can be identified using the knowledge patterns. The degree of predictability is particularly high for verbs identifying concept-related information, especially relations among concepts. For other patterns, the degree of predictability is somewhat smaller since they result in rather different terminological information or noise.

Therefore, I have introduced the concepts of strong and weak knowledge patterns, respectively [9,10]. ‘Strong knowledge patterns’ are patterns with a high degree of proportionality or even constant relations of proportionality with the categories of terminological information. Table 1 shows some important strong knowledge patterns of Danish verbs. One example of a constant relation is the Danish verb *inddele (subdivide)* together with the preposition *i (into).* In this case, the verb phrase will always identify a superordinate concept followed by subordinate concepts as objects in the prepositional construction, as shown in example (1). On the basis of this example, a small concept system can be sketched.

Table 1. Important strong knowledge patterns based on Danish verbs

Terminological information category identified	**Verb + preposition**
superordinate concept + subordinate concepts	*inddele i, opdele i (subdivide into)*
co-ordinate concepts	*skelne mellem (distinguish between)* *adskille sig fra (differ from)*
comprehensive concept + partitive concepts	*bestå af (consist of)* *sammensat af (composed of)*
delimiting characteristics	*karakterisere ved, kendetegne ved (characterize by)*
intensional definitions	*definere som (define as)* *forstå ved (understand by)*

1. Cykler kan **inddeles i** hverdagscykler, sportscykler, transportcykler, HPV-cykler / liggecykler, børnecykler og en lang række andre typer.
   (Bicycles can be ***subdivided into*** *everyday bicycles, sports bicycles, transport bicycles, HPV bicycles, children's bicycles, and a wide range of other types)*

'Weak knowledge patterns', on the contrary, are patterns that result in a high degree of noise, or patterns that result in findings with different types of terminological information requiring a lot of manual work on the part of the terminologist. An important example of the latter is the Danish verb *kalde (call).* Searching on this verb, one may identify terminological information such as terms, synonyms, relations among superordinate and subordinate concepts, and definitions or explanations. This has inspired me to investigate whether for the verb *kalde* (call), recurrent patterns exist over and above its valency pattern proper, i.e. patterns that might support a more precise identification of terminological categories. The study showed that hedges such as *også (also), ofte (often), almindeligvis (usually), tidligere (earlier), i dag (today),* and *undertiden (sometimes)* often co-occur with *kalde (call).* They turned out to be useful in validating the status of a particular term, i.e. whether it should be a synonym, a preferred term, or an obsolete term. In this way, a knowledge pattern such as *kalde (call)* combined with hedges also becomes a strong knowledge pattern for the relation between terms.

Applying knowledge patterns for terminology investigations, my earlier tests showed that a subdivision of the inventory into strong and weak knowledge patterns was advisable and also that the best search strategy was to begin by searching on strong patterns, which made it possible to predict which information categories would be the result [9,10]. For the testing of web-crawled corpora, the same strategy can be recommended. Similar results have been reached independently by Caroline Barrière [2].

Thus, the strong knowledge patterns of verbs have been applied for testing whether a web-crawled corpus has sufficient conceptual coverage. The approach may be criticized because individual knowledge patterns, not specific concepts, are the point of departure. Thus, no overall view of the knowledge patterns occurring together with a given concept will be obtained and therefore no full picture of the amount of conceptual information of a specific concept in a given corpus will be obtained.

6.2 Word Sketch Used for Identifying Knowledge Patterns

Word sketch for languages provided with high-level resources allows the terminologist to gain such an overview of the number of knowledge patterns and thus of the amount of potential conceptual information that may be related to a specific concept in the corpus. As shown in Fig. 3, word sketch contains information on verbs and related prepositions. In the case of *bicycle*, we find the strong knowledge patterns *classify as* and *class as* labeled ‘pp_obj_as’ in the utmost right-hand column. Besides, in the first column ‘object_of’, the verbs *define* and *classify* occur without a preposition.

7 Conclusion

The study aimed at finding a method for evaluating whether web-crawled corpora could be used for exam assignments in terminology. Focus was on linguistic and conceptual coverage as important criteria.

For languages provided with high-level resources in Sketch Engine, the degree of linguistic coverage can be tested by means of the term extraction function. The word sketch is another important support function allowing the terminologist to obtain an overview of the multi-word term candidates related to a specific term.

The degree of conceptual coverage has been tested by searching the corpus for strong knowledge patterns. The word sketch function has also proved well-suited for terminologists because it provides a quick overview of the knowledge patterns occurring together with specific concepts in a given corpus. Consequently, the functions of Sketch Engine can be usefully combined with knowledge patterns for evaluating web-crawled corpora on a qualitative basis, to make sure that the corpora comprise relevant terminological information in knowledge-rich contexts.

For languages not provided with high-level resources, however, the work process must be based on pragmatic, character string approaches. Searches for generic terms as part of composite terms often contribute to creating a good overview of the degree of linguistic coverage, especially when combined with the reduced frequency list, as shown in Fig. 2. To test the corpus for potential multi-word terms, concordance lists are used. To assess the degree of conceptual coverage, tests using searches for the knowledge patterns as the point of departure have been carried out, followed by manual linking of the concordances to the specific concepts. This approach shows that right from the start, testing a Danish terminological corpus consisting of raw text only and carrying out subsequent terminology work will be a much more labor-intensive manual task than a similar task in a language provided with the high-level resources.

Finally, it turned out that many strong knowledge patterns for Danish identify semantic relations among concepts. Experience also shows that knowledge patterns often occur close to each other, and that the texts chunks in which knowledge patterns occur are often heavily loaded with conceptual information. If corpora contain these types of patterns, preliminary concept systems can be worked out on the basis of them. These are useful points of departure for exam assignments, since the first phases of the terminological process consist in analyzing relations among concepts and working out concept systems which in turn form the basis for drafting good definitions.

References

1. Baisa, Vít. “Sketch Engine for Terminology and Translation, webinar prepared for TermNet.” Accessed January 29, 2016.
   https://www.sketchengine.co.uk/category/news/
2. Barrière, Caroline. “Semi-automatic corpus construction from informative texts.” In *Lexicography, Terminology and Translation, Text-Based studies in honour of Ingrid Meyer*, edited by Lynne Bowker, 81-92. Ottawa: University of Ottawa Press, 2006.
3. Bowker, Lynne, and Jennifer Pearson. *Working with Specialized Language, A practical guide to using corpora.* London/New York: Routledge, 2002.
4. Daugaard, Jan, and Sabine Kirchmeier-Andersen. „The Odense Valency Dictionary Programme for Verb Coding.” In *Odense Working Papers in Language and Communication No. 8*, edited by Jan Daugaard, 3-35. Odense: Odense University, 1995.
5. Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. “The Sketch Engine: ten years on.” *Lexicography: Journal of ASIALEX, volume 1* (2014): 7–36.
6. Kilgarriff, Adam. “Getting to know your corpus.” In *Proceedings of The 15th International Conference on Text, Speech and Dialogue (TSD), Czech Republic*, edited by Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala, 3–15. Berlin: Springer, 2012.
7. Meyer, Ingrid. “Extracting knowledge-rich contexts for terminography, A conceptual and methodological framework.” In *Recent Advances in Computational Terminology*, edited by Didier Bourigault, Christian Jacquemin, and Marie-Claude L’Homme, 279-301. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2001.
8. Sketch Engine. ”Writing Sketch Grammars.” Accessed February 16, 2016.
   https://www.sketchengine.co.uk/writing-sketch-grammars/
9. Weilgaard Christensen, Lotte. ”Hvordan ord sporer termer og andre terminologiske oplysninger.” In *Nordterm 2005: ORD OG TERMER, NORDTERM 14,* edited by Ágústa Þorbergsdóttir, 83–93. Reykjavík, 2006.
10. Weilgaard Christensen, Lotte. “Valency Patterns of Danish Verbs as Terminological Knowledge Patterns.” In *Workshop Computational and Computer-assisted Terminology, LREC 2004, IV International Conference On Language Resources and Evaluation,* edited by Rute Costa, Lotte Weilgaard, Raquel Silva, and Pierre Auger, 20-23. Lisbon, 2004.