『コーパスに基づく言語学教育研究報告』No.3（2009）

A Brief Tutorial on Using Collocations for Uncovering and Contrasting Meaning Potentials of Lexical Items[1]

Rainer Perkuhn and Holger Keibel
Institute for the German Language (IDS), Mannheim

Abstract

This introductory tutorial describes a strictly corpus-driven[2] approach for uncovering indications for aspects of use of lexical items[3]. These aspects include '(lexical) meaning' in a very broad sense and involve different dimensions, they are established in and emerge from respective discourses. Using data-driven mathematical-statistical methods with minimal (linguistic) premises, a word's usage spectrum is summarized as a *collocation profile*. Self-organizing methods are applied to visualize the complex similarity structure spanned by these profiles. These visualizations point to the typical aspects of a word's use, and to the common and distinctive aspects of any two words.

1. Introduction

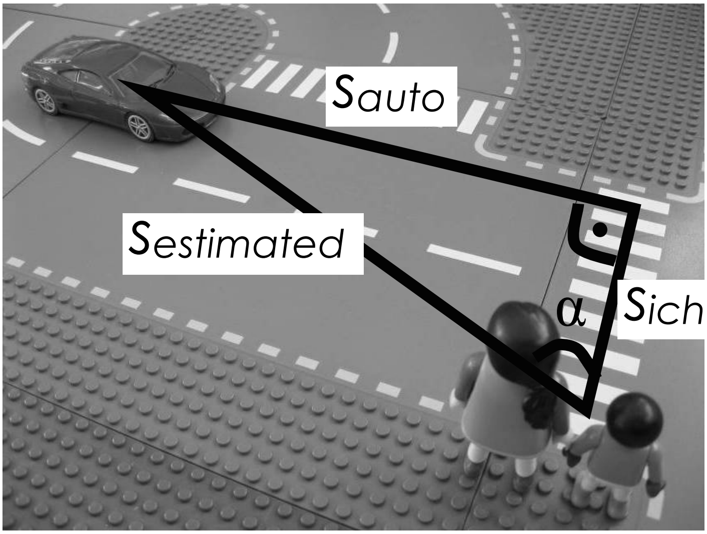
One of the fundamental tenets of the paradigm of 'usage-based linguistics' postulates that linguistic structures emerge from the dynamics in language use (cf. e.g. Bybee 1998; Hopper 1998). The work described in this paper is part of our research programme set up in the spirit of this emergentist perspective (Keibel/Kupietz, this volume). We assume that the fundamental cognitive mechanisms underlying language proficiency rely heavily on tacit knowledge and that they are (in part) functionally equivalent to a statistical assessment of the context (i.e. collocational) behavior of words (or other entities) in the 'language input'. Such a statistical assessment, in turn, depends on some notion of 'frequency' of occurrence of events.

[1] This paper is a summary of two invited talks presented at the Global COE Workshops at TUFS on March, 18/19, 2008, and is based on joint research with Cyril Belica and Marc Kupietz.
[2] For the distinction "corpus-based" vs. "corpus-driven" see Tognini-Bonelli (2001: 84).
[3] For ease of reading we mostly use the familiar (but imprecise) term 'word' instead of the more precise term 'lexical item'. Examples are given in German, but the methodology is language-independent and applicable to any other language as well. Translations are provided where necessary for understanding the phenomena.

According to the emergentist perspective on language, linguistic structures are not fixed but are negotiated in (nearly) every communicative situation. Except maybe for the very early stages of language acquisition these negotiations are primed by the personal experiences about occurrences and recurrences of language use.

Surely, the subjectively experienced “frequency”, i.e. the estimated number of occurrences of events of combinations, is boosted by emotional connotations and declines in time for items that were not recently encountered. In any case, the ‘global’ effects of the dynamics in language use on the emerging phenomena are the result of many interactions between a large number of speakers of a language community. Various factors, including the ambition to be understood in different situations, at different places, at different times, and across generations, result in a certain degree of stableness in language use. The sufficiently stable aspects are generally perceived as the conventions of the language community. In other words, these issues are – partially in form of our cultural heritage – the conserving forces, and have to be considered as impact factors for the dynamics in language use. So, in order to be able to capture the global effects and to reasonably argue for statistical assessments of frequencies as a viable approach, one needs an appropriate empirical basis that incorporates the effects of these intertwined factors. We believe that a very large general-purpose corpus meets these requirements better than any other currently available resource.

Even if the assumptions about the emergent nature of language are true, one still has to cope with the fact that language knowledge is mostly tacit knowledge which cannot be elicited by introspection or by directly asking a native speaker. But a reasonable way could be to design a *functional model* and to demonstrate that the model adequately describes the behavior. As an analogy, everybody crossing a road uses their experience to assess distances and speeds to perform the task successfully (i.e., to avoid accidents) but describing this capability involves a fairly complex trigonometric model. In both scenarios the agents cannot put their knowledge into words or a mathematical formula. But any appropriate model offers a plausible explanation why some agent starts crossing the street or rather waits for the approaching car, or why some speaker uses one expression or another, respectively.



$$s_{auto} = v_{auto} \cdot t_{auto}$$

$$s_{ich} = v_{ich} \cdot t_{ich}$$

$$\tan(\alpha) = \frac{s_{auto}}{s_{ich}}$$

$$\tan(\alpha) = \frac{v_{auto} \cdot t_{auto}}{v_{ich} \cdot t_{ich}}$$

$$\tan(\alpha) \cdot \frac{v_{ich}}{v_{auto}} = \frac{t_{auto}}{t_{ich}}$$

$$t_{ich} < t_{auto} \Leftrightarrow 1 < \frac{t_{auto}}{t_{ich}}$$

$$1 < \tan(\alpha) \cdot \frac{v_{ich}}{v_{auto}}$$

Figure 1: Illustration of how a mathematical model can describe tacit knowledge underlying human behavior

2. Methodology

Our methodology is based on three pillars:

- appropriate data in electronic form (i.e. very large corpora),
- sophisticated statistical/mathematical corpus-linguistic methods,
- the human mind of native or near-native speakers to interpret the analytical results.

It starts from as few linguistic premises as possible and is thus compliant with Sinclair's 'minimum assumption' principle (Sinclair, 1991). Explicitly avoiding a-priori-models of language, the method strictly distinguishes between

- the recorded observation,
- any text-external annotations (regional, diachronic, …), and
- the postponed intellectual interpretation.

In contrast to approaches in computational linguistics and especially in natural language processing, it is not necessarily intended that the outcomes can be operationalized.

As an empirical basis we use the German Reference Corpus DeReKo (cf. Kupietz/Keibel, this volume), a very large corpus archive currently comprising more than 3.4 billion words.

The enormous size of this data collection is the prerequisite for any reasonable application of structure-detecting mathematical methods. All our core methods are based on some form of *correlation analysis* as it is also used in the domain of data mining. Such an analysis assesses whether two given events A and B occur together at a significantly greater frequency than would be expected by mere chance, given the individual number of occurrences of either event. The relevance of a detected correlation, however, can only be decided by a human interpreter.[4]

We use similar methods to detect relations between the occurrences of words or between the occurrences of a word and features of the texts in which it occurs. For example, it is very interesting to investigate the behavior of words over time: new words becoming slowly part of the language vocabulary or old words that have gone out of fashion. In this case, the number of occurrences of words has to be analyzed relative to the feature ‘production date’ or ‘publication date’ of the texts. In many cases, candidates for interesting categories can be recognized by typical patterns in a corresponding time series representation.

3. Collocations and collocation profiles

A very important feature of a text are the words occurring in the lexical context of a word. As Firth (1968:179) noted[5], a word can only be characterized with respect to the lexical contexts in which it occurs. Strictly speaking, the ‘meaning’ of a word cannot be determined without a context. If we argue that each context might contribute to the set of all possible meanings of the word, then an isolated word without context does not have any fully-specified meaning but rather a *meaning potential*. To avoid being misinterpreted we use the term ‘meaning’ with quotation marks in a very broad sense that covers many aspects of a word’s use including connotation and style. In order to understand our approach it might be useful to forget any known framework for the time being (even if there is an affinity to prototypical concepts) and to have an intuitive notion of ‘meaning’ in mind. In a certain way some of the aspects of a word’s use might be viewed as constituting some notion of a ‘word sense’. But again, we doubt that the lexical meaning of a word is adequately captured by listing a few word senses and assuming sharp boundaries between them (cf. Kilgariff, 1996). Instead, we are convinced that a much more fine-grained structure is needed in order to express the meaning potential.

Possibly*,* each single context of a lexical item can contribute to its meaning potential. But note that from the corpus-linguistic point of view, statements must hold even if any given event is ignored. In the same manner as it is possible that a specific context contributes to one or multiple patterns, i.e. as it expresses (different) recurrent aspects of the word’s use, there might be contexts

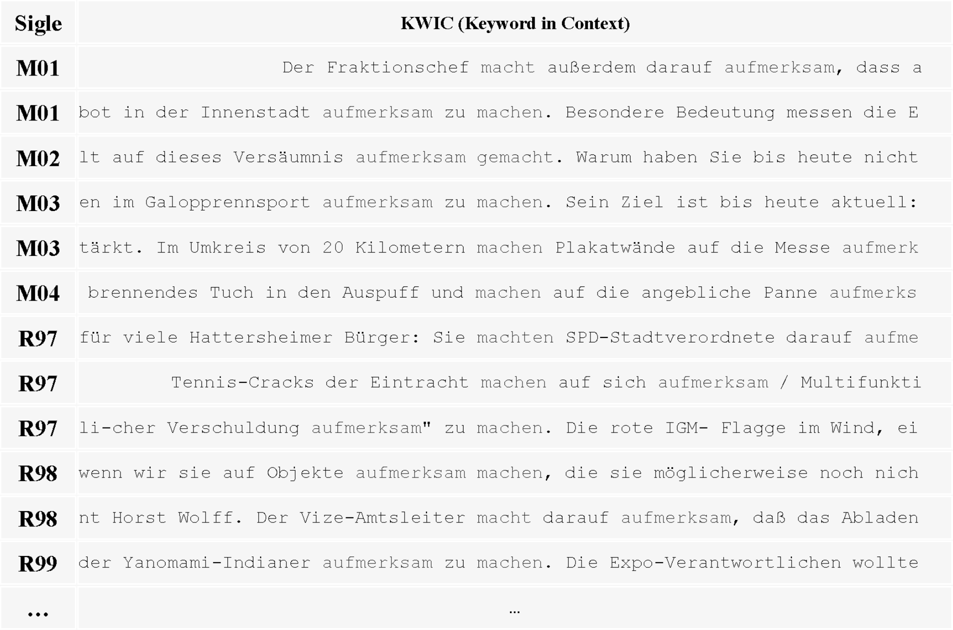
[4] In the domain of data mining, a simple scenario is the market basket analysis which allows to draw conclusions from a large set of contents of market baskets, and to detect what kind of products consumers systematically buy together. This information might be very useful for marketing purposes, or simply for placing the products in the shelves. While the correlations can be detected automatically, an explanation for why two given products are bought together can only be identified by human analyzers. Perhaps the products are ingredients of a single cooking recipe (e.g., flour and sugar) or belong together in another way (such as toothbrush and toothpaste), or are just special offers in the latest leaflet.

[5] “You shall know a word by the company it keeps.”

that do not contribute to any significant pattern. Of course, it might be an important information that (and how often) the lexical item was used in these contexts, but, given the emerging nature of ‘meaning’, especially contexts that express a recurrent aspect of the word’s use are relevant for further investigation.

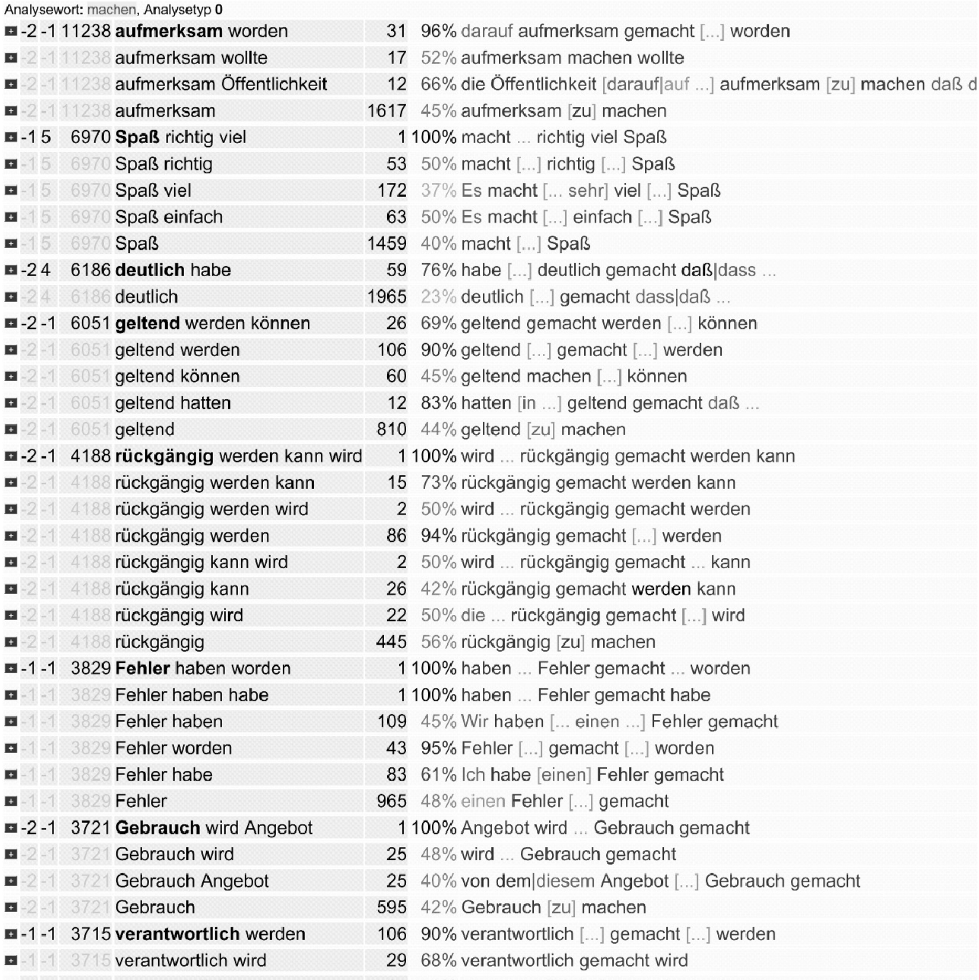
Very simple approaches equate the notion of recurrence with the most frequent *n-grams* (i.e., contiguous word sequences). More sophisticated methods go beyond mere frequency of occurrence and estimate the probability that this frequency might be explained by mere chance. The significant n-grams identified by such an analysis are generally called *(contiguous) collocations*. We use an iterative extension to the standard collocation algorithm (Belica 1995) to extract so-called *higher-order collocations* which are potentially non-contiguous and may occur with flexible word order and word distances. This property is particularly relevant for languages with flexible word order such as German. With this method we can uncover *significant regularities* in the *use of word combinations.*

Applying this method yields many significant higher-order collocations around a given word. Each word combination is represented as a *collocation cluster* comprising the set of all texts in which the word combination occurs (in the form of snippets, either shown as one line or paragraphs, cf. Figure 2) and as the list of words that the word combination consists of.



Sigle	KWIC (Keyword in Context)
M01	Der Fraktionschef macht außerdem darauf aufmerksam, dass a
M01	bot in der Innenstadt aufmerksam zu machen. Besondere Bedeutung messen die E
M02	lt auf dieses Versäumnis aufmerksam gemacht. Warum haben Sie bis heute nicht
M03	en im Galopprennsport aufmerksam zu machen. Sein Ziel ist bis heute aktuell:
M03	tärkt. Im Umkreis von 20 Kilometern machen Plakatwände auf die Messe aufmerk
M04	brennendes Tuch in den Auspuff und machen auf die angebliche Panne aufmerks
R97	für viele Hattersheimer Bürger: Sie machten SPD-Stadtverordnete darauf aufme
R97	Tennis-Cracks der Eintracht machen auf sich aufmerksam / Multifunkti
R97	li-cher Verschuldung aufmerksam" zu machen. Die rote IGM- Flagge im Wind, ei
R98	wenn wir sie auf Objekte aufmerksam machen, die sie möglicherweise noch nich
R98	nt Horst Wolff. Der Vize-Amtsleiter macht darauf aufmerksam, daß das Abladen
R99	der Yanomami-Indianer aufmerksam zu machen. Die Expo-Verantwortlichen wollte
...	…

Figure 2: Some lines of text that constitute the collocation cluster labelled “aufmerksam machen” (Engl. “to call attention”)



Analysewort: machen, Analysetyp **0**

⊞	-2	-1	11238	**aufmerksam** worden	31	96%	darauf aufmerksam gemacht [...] worden
⊞	-2	-1	11238	aufmerksam wollte	17	52%	aufmerksam machen wollte
⊞	-2	-1	11238	aufmerksam Öffentlichkeit	12	66%	die Öffentlichkeit [darauf|auf ...] aufmerksam [zu] machen daß d
⊞	-2	-1	11238	aufmerksam	1617	45%	aufmerksam [zu] machen
⊞	-1	5	6970	**Spaß** richtig viel	1	100%	macht ... richtig viel Spaß
⊞	-1	5	6970	Spaß richtig	53	50%	macht [...] richtig [...] Spaß
⊞	-1	5	6970	Spaß viel	172	37%	Es macht [... sehr] viel [...] Spaß
⊞	-1	5	6970	Spaß einfach	63	50%	Es macht [...] einfach [...] Spaß
⊞	-1	5	6970	Spaß	1459	40%	macht [...] Spaß
⊞	-2	4	6186	**deutlich** habe	59	76%	habe [...] deutlich gemacht daß|dass ...
⊞	-2	4	6186	deutlich	1965	23%	deutlich [...] gemacht dass|daß ...
⊞	-2	-1	6051	**geltend** werden können	26	69%	geltend gemacht werden [...] können
⊞	-2	-1	6051	geltend werden	106	90%	geltend [...] gemacht [...] werden
⊞	-2	-1	6051	geltend können	60	45%	geltend machen [...] können
⊞	-2	-1	6051	geltend hatten	12	83%	hatten [in ...] geltend gemacht daß ...
⊞	-2	-1	6051	geltend	810	44%	geltend [zu] machen
⊞	-2	-1	4188	**rückgängig** werden kann wird	1	100%	wird ... rückgängig gemacht werden kann
⊞	-2	-1	4188	rückgängig werden kann	15	73%	rückgängig gemacht werden kann
⊞	-2	-1	4188	rückgängig werden wird	2	50%	wird ... rückgängig gemacht werden
⊞	-2	-1	4188	rückgängig werden	86	94%	rückgängig gemacht [...] werden
⊞	-2	-1	4188	rückgängig kann wird	2	50%	wird ... rückgängig gemacht ... kann
⊞	-2	-1	4188	rückgängig kann	26	42%	rückgängig gemacht werden kann
⊞	-2	-1	4188	rückgängig wird	22	50%	die ... rückgängig gemacht [...] wird
⊞	-2	-1	4188	rückgängig	445	56%	rückgängig [zu] machen
⊞	-1	-1	3829	**Fehler** haben worden	1	100%	haben ... Fehler gemacht ... worden
⊞	-1	-1	3829	Fehler haben habe	1	100%	haben ... Fehler gemacht habe
⊞	-1	-1	3829	Fehler haben	109	45%	Wir haben [... einen ...] Fehler gemacht
⊞	-1	-1	3829	Fehler worden	43	95%	Fehler [...] gemacht [...] worden
⊞	-1	-1	3829	Fehler habe	83	61%	Ich habe [einen] Fehler gemacht
⊞	-1	-1	3829	Fehler	965	48%	einen Fehler [...] gemacht
⊞	-2	-1	3721	**Gebrauch** wird Angebot	1	100%	Angebot wird ... Gebrauch gemacht
⊞	-2	-1	3721	Gebrauch wird	25	48%	wird ... Gebrauch gemacht
⊞	-2	-1	3721	Gebrauch Angebot	25	40%	von dem|diesem Angebot [...] Gebrauch gemacht
⊞	-2	-1	3721	Gebrauch	595	42%	Gebrauch [zu] machen
⊞	-1	-1	3715	**verantwortlich** werden	106	90%	verantwortlich [...] gemacht [...] werden
⊞	-1	-1	3715	verantwortlich wird	29	68%	verantwortlich gemacht wird

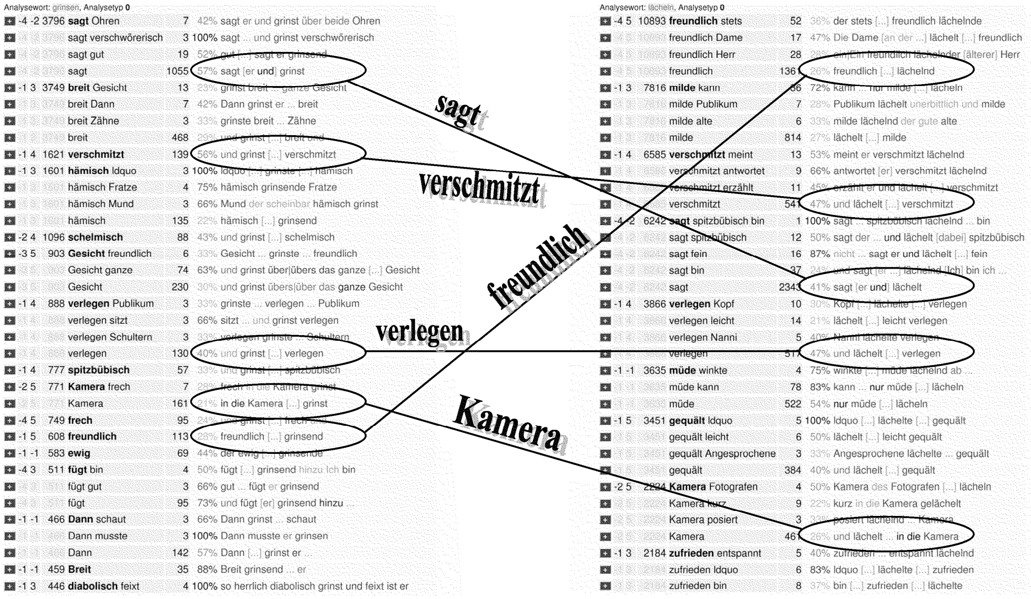
Figure 3: A fragment of the collocation profile of the word "machen" (Engl. "to make"), each line standing for a collocation cluster

The serial word order in the list (of elements of the word combination – one per line in the fourth column in Figure 3) might be counter-intuitive because it does not resemble the order in which the words typically occur in the texts. The predominant word order of a collocation is given in the corresponding *syntagmatic pattern* (cf. Belica, 2003; right-most column in Figure 3) which, in order to improve legibility, also contains wild-card symbols and inserted filler words that were observed to occur in this position at a certain rate.

The set of all collocation clusters above a certain level of statistical significance is called the 'collocation profile' of the given word. If one collocation cluster describes one aspect of the typical use of the word under investigation, then the collocation profile describes the set of the most typical (and presumably most relevant) aspects of its use. We consider the collocation profile a characterization of a word because it captures a large spectrum of aspects or nuances of the word's use, i.e. typical characteristics of the discourse, in which its 'meaning' was established.

4. Uncovering 'meaning'

The notion of 'meaning' has a long tradition in linguistics. Besides relating the 'meaning' of a word to the real world a very broad field in semantics is concerned with meaning relations between words, such as antonymy, hyponomy/hyperonymy, and synonymy. Intuitively, these relations can be characterized by the contexts in which the words are used: For instance, two words are synonymous ('have the same meaning') if they can be interchanged in every context they are used without changing the overall propositional content. If we reformulate this statement by replacing the phrase 'every context' by 'the most typical contexts', we can characterize (near) synonymy in terms of collocation profiles: Two words are likely to be (near) synonyms if they have very similar collocation profiles. In a similar fashion, other types of meaning relations may be characterized in terms of typical contexts that they do or do not have in common. In all these cases, two or more collocation profiles have to be compared. This is a complex task, for a formal comparison has to take into account different sizes of profiles, different positions of the collocates inside the profile and other differences in the quantitative measures of the collocation analysis.



Analysewort: grinsen, Analysetyp **0**

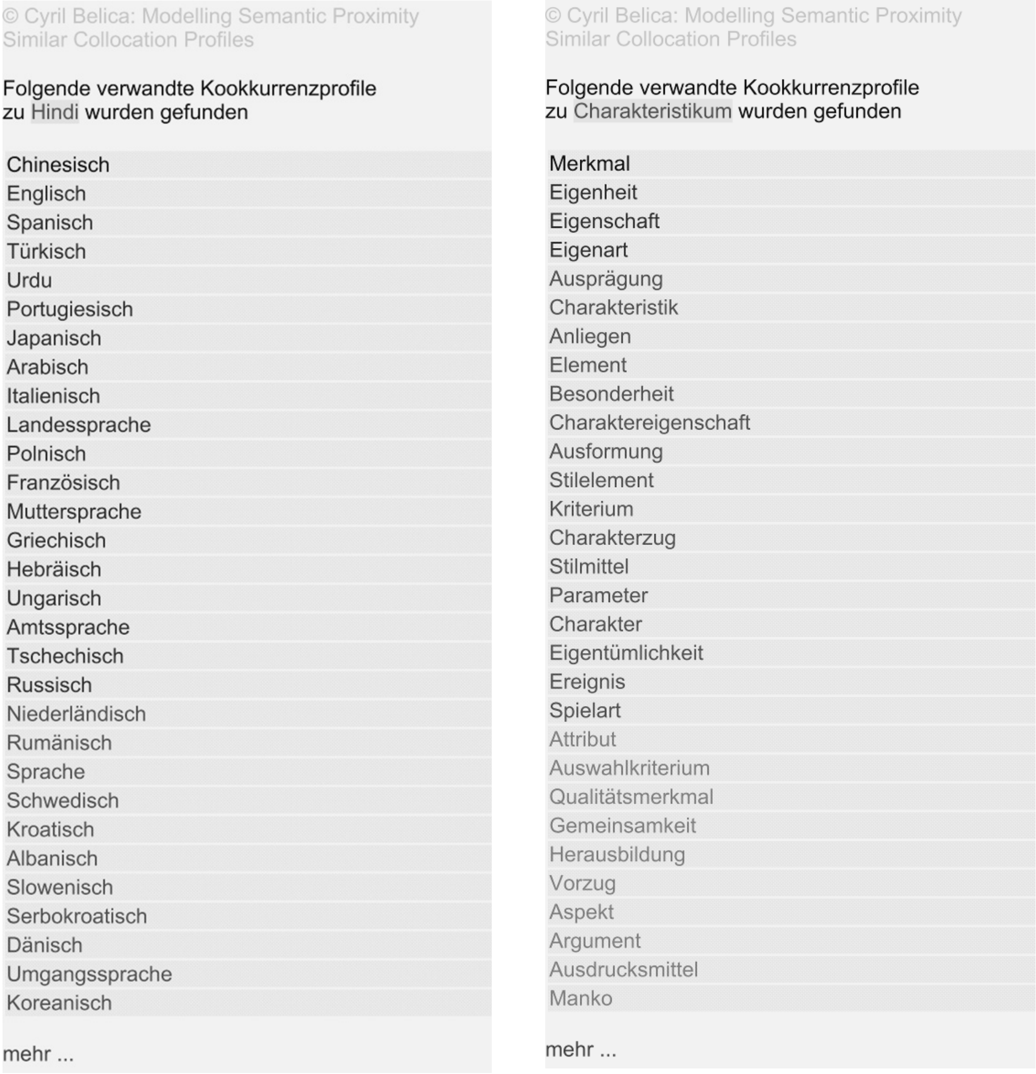
-4 -2	3796	**sagt** Ohren	7	42%	sagt er und grinst über beide Ohren
		sagt verschwörerisch	3	100%	sagt ... und grinst verschwörerisch
		sagt gut	19	52%	gut [...] sagt er grinsend
		sagt	1055	57%	sagt [er und] grinst
-1 3	3749	**breit** Gesicht	13	23%	grinst breit ... ganze Gesicht
		breit Dann	7	42%	Dann grinst er ... breit
		breit Zähne	3	33%	grinste breit ... Zähne
		breit	468	29%	und grinst [...] breit und
-1 4	1621	**verschmitzt**	139	56%	und grinst [...] verschmitzt
-1 3	1601	**hämisch** Idquo	3	100%	Idquo [...] grinste [...] hämisch
		hämisch Fratze	4	75%	hämisch grinsende Fratze
		hämisch Mund	3	66%	Mund der scheinbar hämisch grinst
		hämisch	135	22%	hämisch [...] grinsend
-2 4	1096	**schelmisch**	88	43%	und grinst [...] schelmisch
-3 5	903	**Gesicht** freundlich	6	33%	Gesicht ... grinste ... freundlich
		Gesicht ganze	74	63%	und grinst über|übers das ganze [...] Gesicht
		Gesicht	230	30%	und grinst übers|über das ganze Gesicht
-1 4	888	**verlegen** Publikum	3	33%	grinste ... verlegen ... Publikum
		verlegen sitzt	3	66%	sitzt ... und grinst verlegen
		verlegen Schultern	3	33%	verlegen grinste ... Schultern
		verlegen	130	40%	und grinst [...] verlegen
-1 4	777	**spitzbübisch**	57	33%	und grinst [...] spitzbübisch
-2 5	771	**Kamera** frech	7	28%	frech in die Kamera grinst
		Kamera	161	21%	in die Kamera [...] grinst
-4 5	749	**frech**	95	24%	und grinst [...] frech und
-1 5	608	**freundlich**	113	28%	freundlich [...] grinsend
-1 -1	583	**ewig**	69	44%	der ewig [...] grinsende
-4 3	511	**fügt** bin	4	50%	fügt [...] grinsend hinzu Ich bin
		fügt gut	3	66%	gut ... fügt er grinsend
		fügt	95	73%	und fügt [er] grinsend hinzu ...
-1 -1	466	**Dann** schaut	3	66%	Dann grinst ... schaut
		Dann musste	3	100%	Dann musste er grinsen
		Dann	142	57%	Dann [...] grinst er ...
-1 -1	459	**Breit**	35	88%	Breit grinsend ... er
-1 3	446	**diabolisch** feixt	4	100%	so herrlich diabolisch grinst und feixt ist er

Analysewort: lächeln, Analysetyp **0**

-4 5	10893	**freundlich** stets	52	36%	der stets [...] freundlich lächelnde
		freundlich Dame	17	47%	Die Dame [an der ...] lächelt [...] freundlich
		freundlich Herr	28	28%	ein|Ein freundlich lächelnder [älterer] Herr
		freundlich	1361	26%	freundlich [...] lächelnd
-1 3	7816	**milde** kann	86	72%	kann ... nur milde [...] lächeln
		milde Publikum	7	28%	Publikum lächelt unerbittlich und milde
		milde alte	6	33%	milde lächelnd der gute alte
		milde	814	27%	lächelt [...] milde
-1 4	6585	**verschmitzt** meint	13	53%	meint er verschmitzt lächelnd
		verschmitzt antwortet	9	66%	antwortet [er] verschmitzt lächelnd
		verschmitzt erzählt	11	45%	erzählt er und lächelt [...] verschmitzt
		verschmitzt	541	47%	und lächelt [...] verschmitzt
-4 -2	6242	**sagt** spitzbübisch bin	1	100%	sagt ... spitzbübisch lächelnd ... bin
		sagt spitzbübisch	12	50%	sagt der ... und lächelt [dabei] spitzbübisch
		sagt fein	16	87%	nicht ... sagt er und lächelt [...] fein
		sagt bin	37	24%	und sagt [er ...] lächelnd [Ich] bin ich ...
		sagt	2343	41%	sagt [er und] lächelt
-1 4	3866	**verlegen** Kopf	10	30%	Kopf [...] lächelte [...] verlegen
		verlegen leicht	14	21%	lächelt [...] leicht verlegen
		verlegen Nanni	5	40%	Nanni lächelte verlegen
		verlegen	517	47%	und lächelt [...] verlegen
-1 -1	3635	**müde** winkte	4	75%	winkte [...] müde lächelnd ab ...
		müde kann	78	83%	kann ... nur müde [...] lächeln
		müde	522	54%	nur müde [...] lächeln
-1 5	3451	**gequält** Idquo	5	100%	Idquo [...] lächelte [...] gequält
		gequält leicht	6	50%	lächelt [...] leicht gequält
		gequält Angesprochene	3	33%	Angesprochene lächelte ... gequält
		gequält	384	40%	und lächelt [...] gequält
-2 5	2224	**Kamera** Fotografen	4	50%	Kamera des Fotografen [...] lächeln
		Kamera kurz	9	22%	kurz in die Kamera gelächelt
		Kamera posiert	3	33%	posiert lächelnd ... Kamera
		Kamera	461	26%	und lächelt ... in die Kamera
-1 3	2184	**zufrieden** entspannt	5	40%	zufrieden ... entspannt lächelnd
		zufrieden Idquo	6	83%	Idquo [...] lächelte [...] zufrieden
		zufrieden bin	8	37%	bin [...] zufrieden [...] lächelte

Figure 4: Corresponding elements in two overlapping collocation profiles of the words "grinsen" (Engl. "to grin") and "lächeln" (Engl. "to smile") (only top portions shown)

All following are based on a specific formal measure that quantifies the degree of similarity between any two profiles (Belica, 2004a). Moreover, in order to have fast access to the results of the collocation analysis, more than 220.000 collocation profiles were stored in a static database which constitutes the basis for our “transparent lab” CCDB (Belica, 2007; Keibel/Belica, 2007). The intuitive validity of the formal similarity measure can be verified by inspecting for any given word the list of words whose profiles are most similar to that of the given word, in terms of this measure.



© Cyril Belica: Modelling Semantic Proximity
Similar Collocation Profiles

Folgende verwandte Kookkurrenzprofile
zu Hindi wurden gefunden

Chinesisch
Englisch
Spanisch
Türkisch
Urdu
Portugiesisch
Japanisch
Arabisch
Italienisch
Landessprache
Polnisch
Französisch
Muttersprache
Griechisch
Hebräisch
Ungarisch
Amtssprache
Tschechisch
Russisch
Niederländisch
Rumänisch
Sprache
Schwedisch
Kroatisch
Albanisch
Slowenisch
Serbokroatisch
Dänisch
Umgangssprache
Koreanisch

mehr ...

© Cyril Belica: Modelling Semantic Proximity
Similar Collocation Profiles

Folgende verwandte Kookkurrenzprofile
zu Charakteristikum wurden gefunden

Merkmal
Eigenheit
Eigenschaft
Eigenart
Ausprägung
Charakteristik
Anliegen
Element
Besonderheit
Charaktereigenschaft
Ausformung
Stilelement
Kriterium
Charakterzug
Stilmittel
Parameter
Charakter
Eigentümlichkeit
Ereignis
Spielart
Attribut
Auswahlkriterium
Qualitätsmerkmal
Gemeinsamkeit
Herausbildung
Vorzug
Aspekt
Argument
Ausdrucksmittel
Manko

mehr ...

Figure 5: Two examples of lists of similar profiles

The left-hand part of Figure 5 shows a result that is intuitively plausible purely for semasiological reasons[6] whereas the right-hand part of Figure 5 demonstrates that onomasiological relations[7] may play a role in determining the similarity.

If the collocation clusters in a profile are treated as 'features' of the word nearly all comparisons between profiles result in some common and some distinctive features. But note that, when comparing one word to two different words yields roughly the same similarity values, the common features are not necessarily the same. For instance, the collocation profiles of the words 'mouse' and 'rat' overlap to a high degree. The same is true for the word pair 'mouse' and 'keyboard'. But the overlapping portions are almost completely different in the two cases: In the first comparison, the features overlap that express being an animal, a pet, a parasite, something bred for laboratory experiments (the squares in Figure 6). In the second example, the overlapping features concern the reading of 'mouse' as a computer device (the circles in Figure 6). So, next, we pursue a way to make the resulting partition of the features overt to the interpreter. It will not always be as easy as in the example to label the relevant subsets of the collocation profile by mnemonic descriptions but as a first step we can offer a method to make the distinctions visible.

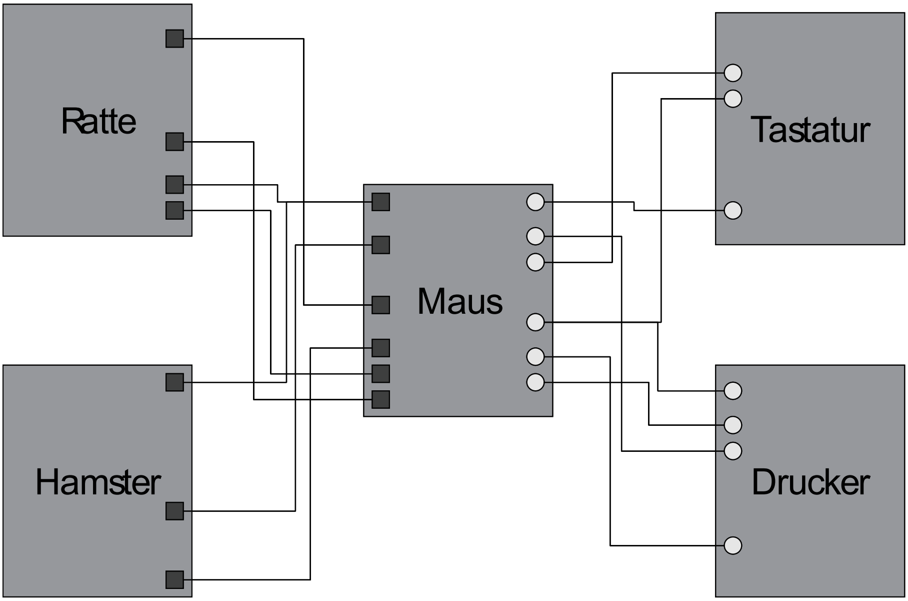


Figure 6: Corresponding aspects in different similarity relations

[6] Au elements of the list for the word 'Hindi' (Engl. Hindi) denote either an individual language or some language domain; most national languages can be recognized from the suffix '-isch' in German.

[7] The elements of the list for 'Charakteristikum' (Engl. feature, characteristic) are all well-motivated but are related to the given word in many different subtle ways.

Provided that a certain aspect of use of a given word is manifested not only in its relation to not only one but multiple other words this should have the effect that the pairwise similarity between any two of these words should be fairly high – due to the shared overlapping aspects – but less similar to most other words.

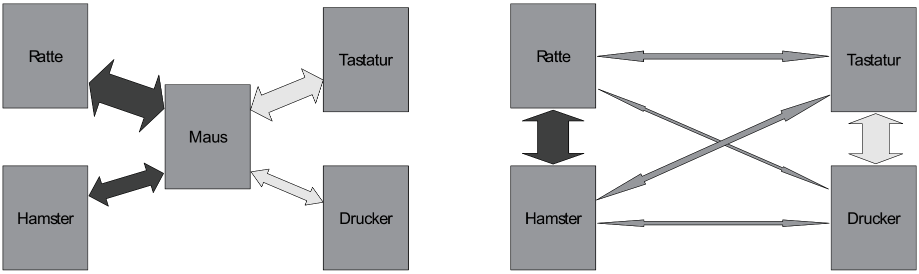


Figure 7: Distinctiveness of aspects (visualized as greyscales) correspond to high similarity measures inside aspect-bearing groups

This prediction is exploited by a second method (Belica 2004b/2005) which takes those words whose collocation profiles are most similar to that of a given word[8] and attempts to cluster them by means of self-organizing feature maps (SOMs) on a colored 5x5 grid[9] so that graphical distance (also of the colors) corresponds to the different degrees of similarity between the groups. The colors of the cells are fixed and depend only on their position. Because the desired distances can be contradictory and not all constraints can be satisfied the procedure iterates starting from different random selections until the representation reaches a nearly stable state. This is why applying the method multiple times generally yields different SOM representations. Nevertheless the different representations typically resemble each other with respect to the overall topography.

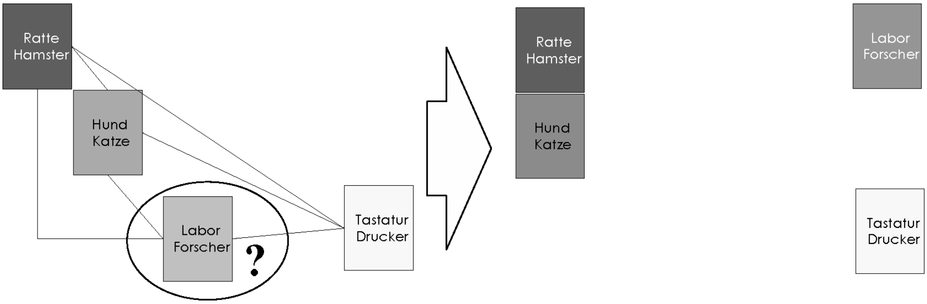
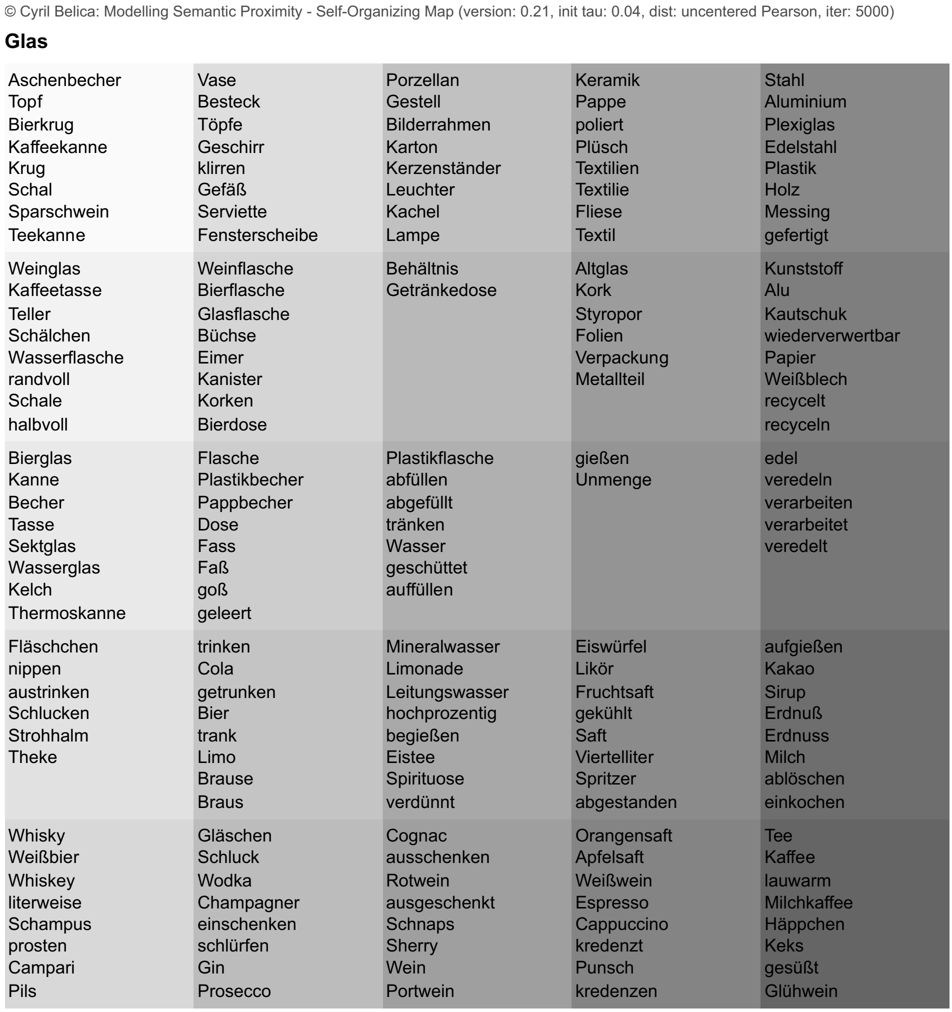


Figure 8: Self-organizing positioning to meet the similarity constraints

[8] Note that the method does not consider the word under investigation (although we should keep in mind that the list of words is defined by the similarity to this word) and only looks at the pairwise relations between all other words of its similarity list.

[9] Due to printing restrictions, the originally colored diagrams are shown here as greyscale figures.



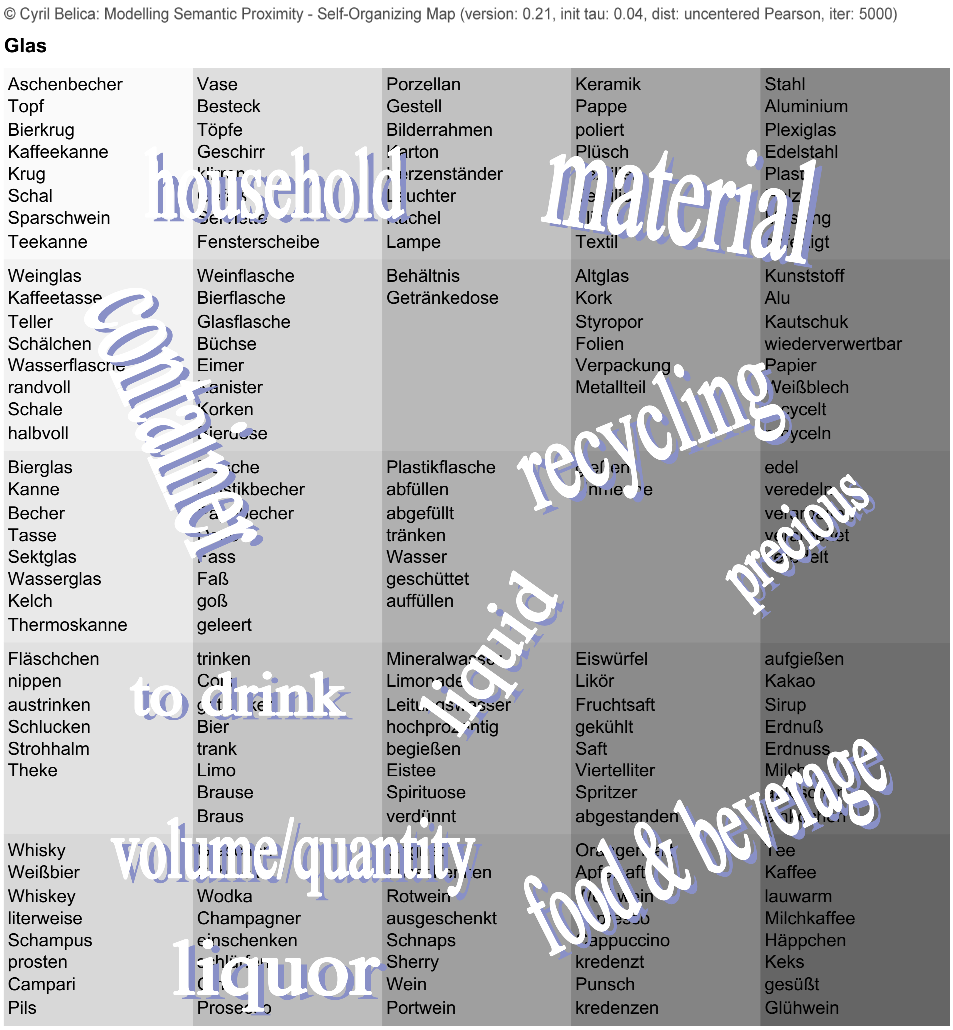
© Cyril Belica: Modelling Semantic Proximity - Self-Organizing Map (version: 0.21, init tau: 0.04, dist: uncentered Pearson, iter: 5000)

Glas

Aschenbecher Topf Bierkrug Kaffeekanne Krug Schal Sparschwein Teekanne	Vase Besteck Töpfe Geschirr klirren Gefäß Serviette Fensterscheibe	Porzellan Gestell Bilderrahmen Karton Kerzenständer Leuchter Kachel Lampe	Keramik Pappe poliert Plüsch Textilien Textilie Fliese Textil	Stahl Aluminium Plexiglas Edelstahl Plastik Holz Messing gefertigt
Weinglas Kaffeetasse Teller Schälchen Wasserflasche randvoll Schale halbvoll	Weinflasche Bierflasche Glasflasche Büchse Eimer Kanister Korken Bierdose	Behältnis Getränkedose	Altglas Kork Styropor Folien Verpackung Metallteil	Kunststoff Alu Kautschuk wiederverwertbar Papier Weißblech recycelt recyceln
Bierglas Kanne Becher Tasse Sektglas Wasserglas Kelch Thermoskanne	Flasche Plastikbecher Pappbecher Dose Fass Faß goß geleert	Plastikflasche abfüllen abgefüllt tränken Wasser geschüttet auffüllen	gießen Unmenge	edel veredeln verarbeiten verarbeitet veredelt
Fläschchen nippen austrinken Schlucken Strohhalm Theke	trinken Cola getrunken Bier trank Limo Brause Braus	Mineralwasser Limonade Leitungswasser hochprozentig begießen Eistee Spirituose verdünnt	Eiswürfel Likör Fruchtsaft gekühlt Saft Viertelliter Spritzer abgestanden	aufgießen Kakao Sirup Erdnuß Erdnuss Milch ablöschen einkochen
Whisky Weißbier Whiskey literweise Schampus prosten Campari Pils	Gläschen Schluck Wodka Champagner einschenken schlürfen Gin Prosecco	Cognac ausschenken Rotwein ausgeschenkt Schnaps Sherry Wein Portwein	Orangensaft Apfelsaft Weißwein Espresso Cappuccino kredenzt Punsch kredenzen	Tee Kaffee lauwarm Milchkaffee Häppchen Keks gesüßt Glühwein

Figure 9: SOM of the word "Glas" (Engl. "glass")

Now it is the task of a human interpreter to examine the resulting SOM. Some preliminary larger-scale studies (Vachkova/Belica to appear) have shown that competent speakers intuitively recognize in a SOM areas of coherent word groups that evoke immediate associations with an aspect of the 'meaning' of the word under investigation and that can be interpreted semiotically. It is generally a good idea to keep track of such interpretations by labeling the identified areas on the SOM accordingly. In a final step, one may attempt to map these findings to the categories of some given linguistic theory – or instead introduce new categories that may be better suited to capture the observed phenomena.



© Cyril Belica: Modelling Semantic Proximity - Self-Organizing Map (version: 0.21, init tau: 0.04, dist: uncentered Pearson, iter: 5000)

Glas

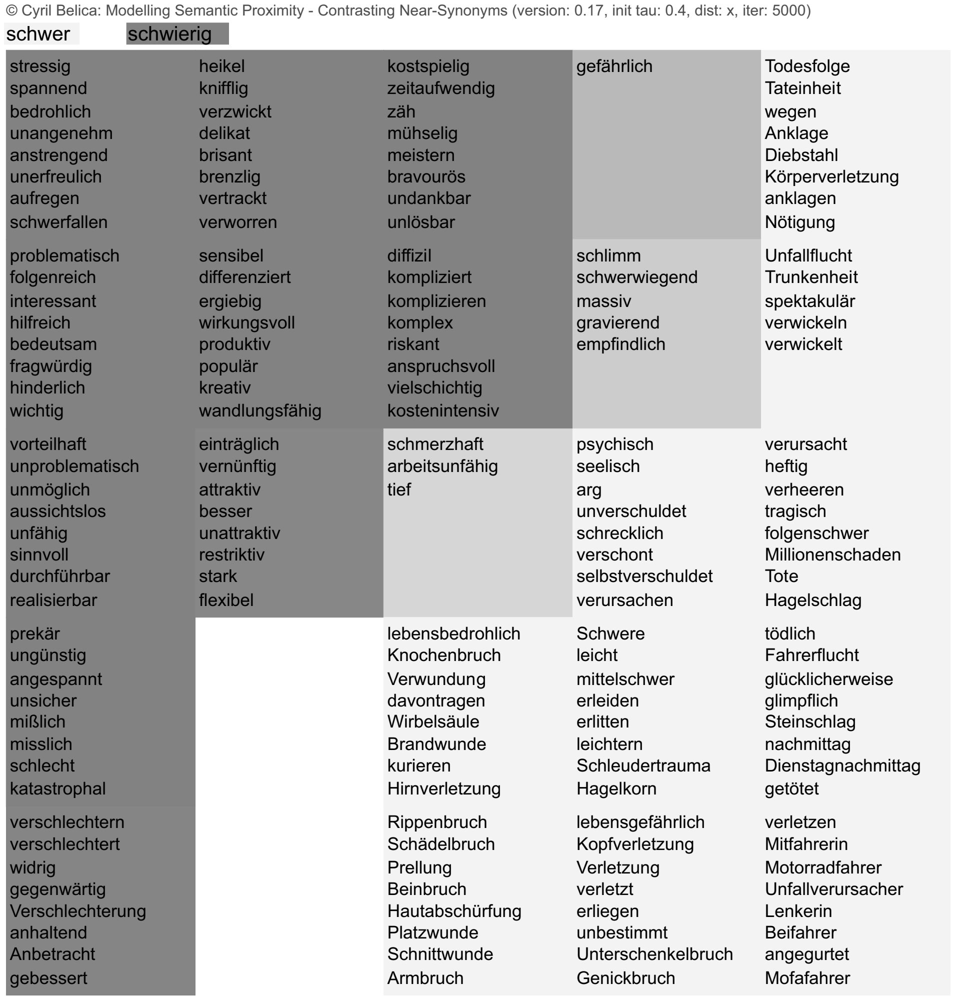
Aschenbecher Topf Bierkrug Kaffeekanne Krug Schal Sparschwein Teekanne	Vase Besteck Töpfe Geschirr klirren [illegible] Serviette Fensterscheibe	Porzellan Gestell Bilderrahmen Karton Kerzenständer Leuchter Kachel Lampe	Keramik Pappe poliert Plüsch [illegible] [illegible] [illegible] Textil	Stahl Aluminium Plexiglas Edelstahl Plast Holz Messing gefertigt
Weinglas Kaffeetasse Teller Schälchen Wasserflasche randvoll Schale halbvoll	Weinflasche Bierflasche Glasflasche Büchse Eimer Kanister Korken Bierdose	Behältnis Getränkedose	Altglas Kork Styropor Folien Verpackung Metallteil	Kunststoff Alu Kautschuk wiederverwertbar Papier Weißblech recycelt recyceln
Bierglas Kanne Becher Tasse Sektglas Wasserglas Kelch Thermoskanne	Flasche Plastikbecher Pappbecher [illegible] Fass Faß goß geleert	Plastikflasche abfüllen abgefüllt tränken Wasser geschüttet auffüllen	[illegible] [illegible]	edel veredeln [illegible] [illegible] [illegible]
Fläschchen nippen austrinken Schlucken Strohhalm Theke	trinken Cola getrunken Bier trank Limo Brause Braus	Mineralwasser Limonade Leitungswasser hochprozentig begießen Eistee Spirituose verdünnt	Eiswürfel Likör Fruchtsaft gekühlt Saft Viertelliter Spritzer abgestanden	aufgießen Kakao Sirup Erdnuß Erdnuss Milch [illegible] einkochen
Whisky Weißbier Whiskey literweise Schampus prosten Campari Pils	[illegible] [illegible] Wodka Champagner einschenken schlürfen [illegible] Prosecco	[illegible] [illegible] Rotwein ausgeschenkt Schnaps Sherry Wein Portwein	Orangensaft Apfelsaft Weißwein Espresso Cappuccino kredenzt Punsch kredenzen	Tee Kaffee lauwarm Milchkaffee Häppchen Keks gesüßt Glühwein

Figure 10: Results of the semiotic interpretation of a SOM

A very similar approach also supports the study of the complex range of contrasts and commonalities in the use of any two words – where near-synonyms are of particular interest. In this case, the same self-organizing method is applied to the set of words whose profiles are most similar to either given word (cf. Belica 2006). In the resulting SOM visualization, the grid is color-marked such that reddish areas contain words that are more similar to the first word and yellowish areas are more similar to the second word while shades of orange model a gradual transition between these two extremes[10]. Inspecting the orange areas in such a SOM, one may identify regions which express the common features of the two given words (i.e. the aspects which contribute to 'meaning

[10] In Figure 11 the 'reddish' is printed as 'dark grey', 'yellowish' shows as 'light grey', and the orange fones appear as greyscales in between.

the same'), and inspecting the clearly reddish or yellowish areas may yield other regions that correspond to distinctive features of either word. Thus, a SOM entirely colored in orange indicates that the two given words are synonymous in virtually all possible contexts.



© Cyril Belica: Modelling Semantic Proximity - Contrasting Near-Synonyms (version: 0.17, init tau: 0.4, dist: x, iter: 5000)

schwer schwierig

stressig spannend bedrohlich unangenehm anstrengend unerfreulich aufregen schwerfallen	heikel knifflig verzwickt delikat brisant brenzlig vertrackt verworren	kostspielig zeitaufwendig zäh mühselig meistern bravourös undankbar unlösbar	gefährlich	Todesfolge Tateinheit wegen Anklage Diebstahl Körperverletzung anklagen Nötigung
problematisch folgenreich interessant hilfreich bedeutsam fragwürdig hinderlich wichtig	sensibel differenziert ergiebig wirkungsvoll produktiv populär kreativ wandlungsfähig	diffizil kompliziert komplizieren komplex riskant anspruchsvoll vielschichtig kostenintensiv	schlimm schwerwiegend massiv gravierend empfindlich	Unfallflucht Trunkenheit spektakulär verwickeln verwickelt
vorteilhaft unproblematisch unmöglich aussichtslos unfähig sinnvoll durchführbar realisierbar	einträglich vernünftig attraktiv besser unattraktiv restriktiv stark flexibel	schmerzhaft arbeitsunfähig tief	psychisch seelisch arg unverschuldet schrecklich verschont selbstverschuldet verursachen	verursacht heftig verheeren tragisch folgenschwer Millionenschaden Tote Hagelschlag
prekär ungünstig angespannt unsicher mißlich misslich schlecht katastrophal		lebensbedrohlich Knochenbruch Verwundung davontragen Wirbelsäule Brandwunde kurieren Hirnverletzung	Schwere leicht mittelschwer erleiden erlitten leichtern Schleudertrauma Hagelkorn	tödlich Fahrerflucht glücklicherweise glimpflich Steinschlag nachmittag Dienstagnachmittag getötet
verschlechtern verschlechtert widrig gegenwärtig Verschlechterung anhaltend Anbetracht gebessert		Rippenbruch Schädelbruch Prellung Beinbruch Hautabschürfung Platzwunde Schnittwunde Armbruch	lebensgefährlich Kopfverletzung Verletzung verletzt erliegen unbestimmt Unterschenkelbruch Genickbruch	verletzen Mitfahrerin Motorradfahrer Unfallverursacher Lenkerin Beifahrer angegurtet Mofafahrer

Figure 11: Contrasting two near-synonyms "schwer" and "schwierig" (Engl. "hard" and "difficult", respectively)

5. Conclusion

Starting from a nearly psychological motivation we have presented a corpus-driven methodology that, given an appropriate empirical basis, allows to uncover meaning aspects – or more precisely: the meaning potential – of a lexical item by detecting its typical contexts (viz., its collocations) and summarizing its most relevant aspects of its use in the form of a collocation profile. Any such profile is considered a point in a high-dimensional space, and self-organizing methods are applied in order to visualize the complex similarity structure in this space around a given profile. This paper has demonstrated how this general methodology can be used to study the ‘meaning potential’ of a given word, or to study the commonalities and contrasts between any two words. This involves intuitive, spontaneous interpretations by competent speakers, potentially followed by a linguistic interpretation. We are confident that studying a large number of words in this way will provide valuable insights into the actual categories underlying the language ‘system’ and its functionality.

References

Belica, Cyril. (1995) *Statistische Kollokationsanalyse und -clustering*, Korpuslinguistische Analysemethode. http://corpora.ids-mannheim.de/.

Belica, Cyril. (2003) *Ermittlung syntagmatischer Ordnungsmuster von Kookkurrenzprofilen*, Korpusanalytische Methode. http://corpora.ids-mannheim.de/ccdb/.

Belica, Cyril. (2004a) *Analyse von Verwandtschaftsrelationen zwischen Kookkurrenzprofilen*, Korpusanalytische Methode. http://corpora.ids-mannheim.de/ccdb/.

Belica, Cyril. (2004b) *Modellierung semantischer Nähe: Verwendungsaspekte. Hierarchische Ähnlichkeitsrelationen zwischen Kookkurrenzprofilen*, Korpusanalytische Methode. http://corpora.idsmannheim.de/ccdb/.

Belica, Cyril. (2005) *Modellierung semantischer Nähe: Analyse und topografische Visualisierung von Verwendungsaspekten in Self-Organizing-Maps*, Korpusanalytische Methode. http://corpora.idsmannheim.de/ccdb/.

Belica, Cyril. (2006) *Modellierung semantischer Nähe: Kontrastierung von nahen Synonymen*, Korpusanalytische Methode. http://corpora.ids-mannheim.de/ccdb/.

Belica, Cyril. (2007) *Kookkurrenzdatenbank CCDB – V3*, Eine korpuslinguistische Denk- und Experimentierplattform für die Erforschung und theoretische Begründung von systemisch-strukturellen Eigenschaften von Kohäsionsrelationen zwischen den Konstituenten des Sprachgebrauchs. http://corpora.ids-mannheim.de/ccdb/.

Bybee, Joan L. (1998) “The emergent lexicon”, *Chicago Linguistic Society* 34, 421-435.

Firth, John R. (1968) “A Synopsis of Linguistic Theory 1930-1955”, *Studies in Linguistic Analysis*, Philological Society, Oxford, 1957. Reprinted in Palmer, F. (ed). *Selected Papers of J.R. Firth*. Harlow: Longman, 168-205.

Hopper, Paul. (1998): “Emergent grammar”, *The new psychology of language. Cognitive and functional approaches to language structure*, Tomasello, Michael (ed)., London: Erlbaum, 155-175.

Keibel, Holger / Belica, Cyril. (2007) “CCDB: A Corpus-Linguistic Research and Development Workbench”, *Proceedings of Corpus Linguistics 2007*, Birmingham.

http://corpus.bham.ac.uk/corplingproceedings07/paper/134_Paper.pdf.

Keibel, Holger / Kupietz, Marc. (this volume) *Approaching Grammar: Towards an empirical linguistic research programme*, this volume.

Kilgarriff, Adam. (1996) "Word senses are not bona fide objects: implications for cognitive science, formal semantics, NLP", *Proceedings of 5th Conference on the Cognitive Science of Natural Language Processsing*, Dublin, 193-200.

Kupietz, Marc / Keibel, Holger. (this volume) *The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research*, this volume.

Tognini-Bonelli, Elena. (2001) *Corpus Linguistics at Work*, *Studies in Corpus Linguistics 6*, Amsterdam: John Benjamins.

Sinclair, John. (1991) *Corpus, Concordance, Collocation*, Oxford.

Vachkova, Marie / Belica, Cyril. (to appear) "Self-Organizing Lexical Feature Maps. Semiotic Interpretation and Possible Application in Lexicography", *Interdisciplinary Journal for Germanic Linguistics and Semiotic Analysis*, *13*, *2*.