Going Online with a German Collocations Dictionary

Tobias Roth

University of Basel, Deutsches Seminar, Nadelberg 4, Basel, Switzerland
tobias.roth@unibas.ch

Abstract

Although a lot of dictionaries are available on the Web, there are no well-established ways to present collocations dictionaries for language learners online. In the online version of the collocations dictionary for German we are working to overcome certain shortcomings of printed collocations dictionaries. A major issue when they are used in production situations (e.g. writing processes) is how to find collocations efficiently. Another difficulty for users is to transfer the information found to their own language use. The lexicographic challenge consists of conceiving a microstructure that assists users in finding a collocation without having to read complete articles. At the same time, enough information has to be given in order for learners to be able to use a collocation appropriately. Our online dictionary uses present-day electronic search facilities for improved access, as well as a presentation of dictionary articles on two levels: a minimalistic view for the search and navigation stage and a more detailed view once a collocation is found.

Keywords: online dictionary, collocations, dictionary design, learners' dictionary, German language

1. Introduction

Many dictionaries are available on the Web today. However, as yet there are no well-established ways of how to present collocations dictionaries for language learners online.

Major issues are retrievability and information transfer. How can a collocation be found efficiently, i.e. without needing to read complete dictionary articles? And how is the information best presented so that users understand the entries and can effectively use collocations found in the dictionary themselves?

The *Kollokationenwörterbuch*[1] – the collocations dictionary we are working on – is not a pure online project. The dictionary is also intended to appear in print. Not all aspects, therefore, are optimised for the online version. Some decisions reflect a compromise between these two mediums as it would not make sense to duplicate certain structures because of slightly differing needs between online and print versions. However, as it is a completely new dictionary built from scratch, it was possible to freely choose the design of the underlying database and special attention

[1] Its full (working) title being Kollokationenwörterbuch – typische und gebräuchliche Wortverbindungen des Deutschen, it is accessible at http://www.kollokationenwoerterbuch.ch.

was paid to the possibility of media-independent publishing.

2. Collocations in online dictionaries

Most learners' dictionaries specifically dedicated to collocations are not available online[2] or are available only in e-book versions as electronic copies of printed dictionaries, such as Quasthoff (2011) for German. However, there are some more general online dictionaries that cover collocations and multi-word units in general. Examples for German are *Duden online*, *elexiko* (Klosa, Schnörch & Storjohann 2006), *DWDS* (Geyken 2011; Klein 2004) and *LEO*.

These dictionaries use different strategies for presenting multi-word units. They range from writing simple listings, separate article entries for each unit and more elaborate visualisation methods, such as word clouds and network graphs.

2.1 Separate entries

In *elexiko*'s sub-dictionary "feste Wortverbindungen", every multi-word unit gets its own article entry. Most have at least slightly idiomatic meanings, so detailed explanations are clearly justified.

For collocations that are semantically compositional, this structure is less appropriate. This would result in very small and relatively uninformative article entries, whereas important aspects such as the context of a collocation (whether there are similar collocations with the same component words) or its retrievability will be neglected.

2.2 Listings

Other dictionaries give simple listings of collocations or multi-word units for a headword. In certain cases they are presented like usage samples, although a majority of these sample combinations possess collocational characteristics and would enter a collocations dictionary (cf. e.g. *Duden online*, the dictionary part of *DWDS*).

Often such listings are not further hierarchically structured. In cases where they are, criteria are often syntactic. For example, the *Wortprofil* (word profile) in *DWDS* groups collocations by their syntactic configuration (there are groups for collocates as subjects, objects, attributes etc.; see also Geyken 2011).

Listings are easy to produce and can potentially display large numbers of collocations in a limited space, but as it becomes more extensive, navigation can become difficult.

[2] *DiCE*, an online collocations dictionary for Spanish, can be cited as one of the few exceptions.

2.3 Word clouds and network graphs

Word clouds and network graphs are more sophisticated tools to visualise collocations of a given headword. Word clouds are used by *Wortprofil* in *DWDS* and network graphs by *Wortschatz Leipzig*; while *Duden online* uses a combination of word clouds and network graphs to display typical word combinations.

Both word clouds and network graphs are preferred for automatically extracted collocation lists. They are hardly ever found in manually crafted articles. The advantage of both of them is a rather compact mode of presentation and the possibility to visualise context and frequencies and the strength of connections.

3. Issues in collocations lexicography

The aforementioned general online dictionaries are obviously not specialised in the presentation of collocations. To obtain a clearer idea about the difficulties one has to deal with in online collocations dictionaries, it is best to start with the analysis of the main issues in collocations lexicography for language learners.

The present project, like many other collocations dictionaries, is perceived to be an aid in text or language production. The prototypical user wants to write or say something about, e.g. a *mountain*, knows that this is *Berg* in German, and expects to find collocations with *Berg* (under the headword *Berg*) that match the meaning he/she wants to convey.

The two main problems here are navigation and information transfer.[3] How should collocations be arranged in the dictionary so they can be retrieved as easily and efficiently as possible? And in what form should the information be provided for users to be able to actually integrate a collocation found in the dictionary into their own speech and writing?

3.1 Retrieval

How collocations are best retrieved is by no means a trivial question. If we consider collocations as transparent and essentially compositional in meaning[4] we can assume that users will be able to look up a collocation under one of its component words. If a headword comprises a large number of collocations the next question is how to group and sort them to ease the search process.

[3] Issues no less important, but more closely related to content, e.g. selection criteria for collocations or integration of compounds (cf. Häcki Buhofer 2011; Roth 2012b), are not discussed here.

[4] As *idioms of encoding* (Fillmore, Kay & O'Connor 1988; Makkai 1972).

3.1.1 Node and collocate vs. base and collocator

Hausmann (1985) introduced the concept of *base* and *collocator* in collocations. The formerly used terms *node* and *collocate* (Sinclair 1966) just indicate a perspective: *node* is the word that is being looked at and its *collocates* are the partner words that form collocations with it. All components of a collocation can be both *node* and *collocate*, just depending on the perspective.

In contrast, *base* and *collocator* describe an absolute hierarchy within a collocation. Rather vaguely defined, the *base* is the word a prototypical user would look up in order to find a collocation; the *collocator* its counterpart. According to Hausmann (1985), the noun is the most important word class for *bases* because nouns denote things and phenomena in the world that we talk about:

> Die wichtigste Basiswortart ist das Substantiv, weil es die Substantive sind, welche die Dinge und Phänomene dieser Welt ausdrücken, über die es etwas zu sagen gibt. Adjektive und Verben kommen als Basiswörter nur insoweit in Frage, als sie durch Adverbien weiter determiniert werden können. (Hausmann 1985, p. 119).

In verb-noun collocations the *base* is the noun; in adjective-noun collocations it is also the noun; in verb-adverb collocations it is the verb, etc. Even if the concept has its problems (cf. e.g. Handl 2009; Herbst 2009; Roth 2012b; Steyer 2000) it has been widely adopted by current collocations dictionaries (Le Fur 2007; Lo Cascio 2012; OCDSE 2009; Quasthoff 2011; Rundell 2010). In printed dictionaries it allows for a reasonable navigation structure without the need of duplicated entries: collocations are printed in the base article only, not under the collocator.

3.1.2 Grouping and sorting

Several proposals have been made on how to arrange collocations within an article. Grouping and sorting criteria are mainly morphosyntactic, syntactic and semantic. As outlined above, the search process on this level is semantically motivated: users look for a collocation that fits, as closely as possible, the meaning they want to express. They might have an idea of how the construction of the whole sentence will appear, hence the morphosyntactic and syntactic criteria, but essentially it is a semantic choice.

Most collocations dictionaries have at least two grouping levels below the headword.[5] Quasthoff (2011) groups according to word class (verb, adjective). Noun-verb collocations are subgrouped according to the grammatical case of the base noun, whereas collocations with adverbs and adjectives contain semantically motivated subgroups. The OCDSE (2009) and Rundell (2010) both consider word class groups on the top level, but include positional information (e.g. *X + verb* vs. *verb + X*). Subgroups are semantically motivated; in the case of Rundell (2010) the content of a

[5] The exact number depends on whether the splitting of different meanings of a headword is considered as a grouping level or not.

semantic subgroup is explicitly stated. Le Fur (2007) also forms groups by word class, positional information and semantically motivated subgroups. Finally, Lo Cascio (2012) forms the same top level groups (word class and positional information), but no sub-groups; instead, the collocations are in alphabetical order.

3.2 Information transfer

Once a suitable collocation is found, a user needs to know its exact form and properties so as to be able to actually use it. In a preliminary study on article structure conducted at different schools, students preferred less abstract citation forms and articles with more example sentences (Siebenhüner 2010). They often displayed difficulties in deriving the correct usage of a collocation from collocators only or from abstract citation forms without examples.

This need to be more explicit in order to facilitate information transfer contradicts in some ways the need to be as compact as possible in order to facilitate navigation and retrieval. Most current collocation dictionaries focus on compactness rather than on explicit information presentation.

The majority provide the base form of the collocator but no citation form more explicit (Le Fur 2007; OCDSE 2009; Quasthoff 2011; Rundell 2010). An exception to this is Lo Cascio (2012) who provides extended citation forms. Others try to convey grammatical information mainly by their structure of groups and subgroups (see above). Le Fur (2007) additionally indicates certain grammatical or other features by means of abbreviations in superscript next to the collocator. Example sentences are given by most of the dictionaries quoted above. Exceptions are Quasthoff (2011) who gives no example sentences at all and Lo Cascio (2012) with explicit meaning indications for a big part of the collocations.

4. A German online collocations dictionary

The present project consists of creating a German collocations dictionary with collocations of about 2000 base-vocabulary headwords (Häcki Buhofer 2011; Roth 2012b). The primary target audience is intermediate L2 learners of German. The dictionary is not an online-only project; there will also be a printed version.

The dictionary is completely new, written from scratch, so there was no need to consider the integration of older versions or other kinds of legacy data. The dictionary writing system in use has also been newly developed for this specific purpose. This offered the possibility to structure the data in such a way that would allow media-independent publication (Roth 2012b). Online and printed presentations are not completely independent, however, as they share certain common features. On one hand they share the same needs concerning some points, whereas on the other hand it would often be highly uneconomical to duplicate features with only slight differences between online and printed versions. Sometimes

there is a solution that is suitable for both versions, even if there was a more ideal solution for a particular format, and in such cases a common approach is utilised for both.

A prototype of the online version of the *Kollokationenwörterbuch* can be found under the URL http://www.kollokationenwoerterbuch.ch. Its main characteristics, and some proposals for solutions to the presentation issues raised above, are described below.

4.1 Search

The main means of interaction with the dictionary is a simple search field, similar to the familiar Web search engine types. When typing a word into the field, matching lemmas show up in a menu list underneath in an 'as-you-type' fashion (see Figure 1).

At the top of the list there are words beginning with the search term, whereas below you can find words containing the search term. Lemmas that are part of the 2000-item base vocabulary appear in bold. For these lemmas a complete collocation search including manual semantic grouping (see 4.2.) has been performed. Lemmas not in bold appear in collected collocations, but they have not been treated as a headword for the printed version and they have not undergone a systematic collocation search. These articles are dynamically assembled. As no manual semantic grouping has taken place in these cases, collocations are presented alphabetically, grouped by word class.

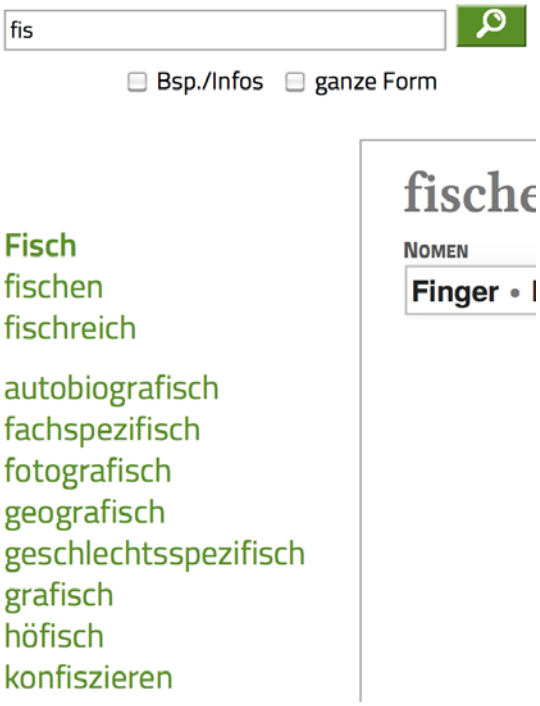

Figure 1: Search and navigation

If you type more than one word in the search field, the dictionary article for the first search word is fetched and subsequent search strings are highlighted in the just-loaded

article (see Figure 2-c).

Such standard search functionality for an online application helps in overcoming the problem of whether it is reliably the base that is looked up. Access through collocators is possible, also, and all collocations belonging to a word are directly shown. Articles do not strictly follow the base-collocator principle anymore, but rather show a node-collocate approach. Yet, the overall structure of an article is not greatly changed because of this. What would otherwise appear as links to other articles are now presented as full collocation entries, but displayed in a separate grouping at the end of the article.

In general, the possibility to easily search by all collocation components is a big improvement in retrievability.

4.2 Grouping and navigation layer

Once one component word (*node*) is found along with its associated collocations, the next question or challenge is how to find a suitable collocation without having to read the complete article.

In the present project, it was decided to introduce two hierarchical grouping levels. In a first step, collocations are grouped by the word class of their collocates (see Figure 2 b). Subsequently, they are further subgrouped according to semantic criteria. Collocations belonging semantically together can be found in the same subgroup. A subgroup may receive a label (see Figure 2e) that describes its content or is at least associated with the collocates of this subgroup and stands as a kind of a prototypical example. Its goal is not the meaning description proper, but to assist in navigation.[6]

This also holds for the printed version. The main difference introduced in the online dictionary is a split into two presentation layers (see Figure 2). On the first layer, still in the navigation stage, only collocates are displayed. All supplementary information, such as extended citation forms, example sentences, meaning indications, etc., is omitted. The collocates are displayed in boxes grouped by semantic similarity. With this layout, a maximum of collocations fit on one screen in a clearly arranged fashion. This should help users to more quickly find the collocations they seek. With only one word per collocation a strict minimum of information is provided with no extra information to detract from the retrieval task.

4.3 Detailed information

The second layer presents all collected information for a collocation. As soon as a suitable collocation is found the one-word-per-collocation approach has reached its goal and is then no longer informative enough. The user's next task is to find out how

[6] Not like in Rundell (2010) where actual meaning descriptions for every subgroup are given.

exactly to use this collocation. Studies conducted in this project (Siebenhüner 2010) have confirmed that extended citation forms and a large number of example sentences are necessary for many students so that they can correctly use collocations they have looked up.

Along with extended citation forms and example sentences some additional detailed information is presented here. Some collocations that might be difficult to understand are given meaning indications. Pragmatical usage information is also provided (markers such as *informal*, *pejorative*, etc., but also more detailed usage explanations when considered necessary). Collocations are also marked for regional usage restrictions on a country level for Austria, Germany and Switzerland (Roth 2012a).

This detailed information on the second layer is accessed by a clicking or hovering action on the single collocates causing an expansion of the details window (Figure 2d). In addition, this approach has the advantage that users interact with the dictionary application with more actively involvement than just by plain reading.

4.4 Internal and external links

The detailed view of a collocation provides links to internal and external targets. For the time being this feature is not extensively used; so far there are internal links to other dictionary articles and external corpus links.

Internally, collocates in the detailed view are linked to the corresponding node articles. Clicking on a collocate will open an article with the respective word as a node. If, in the example article in Figure 2, the user is unsure of the exact meaning of *ausrücken* they can click on it and navigate to the article for *ausrücken*. There, the user will find collocations that inform that the word describes something that the fire brigade (*Feuerwehr*) and the police (*Polizei*) do. If the user required a collocation describing the arrival of the fire brigade they can now click on *Feuerwehr* to obtain several verbs that can be used for this purpose.

The first external links given have the *Swiss Text Corpus* (Bickel et al., 2009) as a target. These links will open the corpus site with a *KWIC* view (key word in context) of examples for the respective collocation.

Besides links to more corpora, links to other dictionaries could also prove useful. Collocations dictionaries do not provide information about individual words, such as meaning indications, grammatical information beyond citation forms and examples as well as pragmatical usage information, which can be seen as a shortcoming. Links from individual words to a general dictionary or even to a bilingual dictionary could help users in this case (but do not form part of the current version of the *Kollokationenwörterbuch*).

4.5 Customisability

An advantage of online dictionaries and electronic dictionaries in general is that they are more dynamic. Potentially, everyone can have their own, tailor-made version of a dictionary in terms of what data are displayed and how they are displayed.

However, users also expect an online dictionary to work 'out-of-the-box'. Their first concern will typically not be how they can customise it. In addition, it is often not very clear what special features users might expect from an online dictionary, as Müller-Spitzer, Koplenig & Töpel (2011) put it:

> Nevertheless, this does not mean that the development of innovative features of online dictionaries is pointless. As we show elsewhere in detail [...], users tend to appreciate good ideas, such as a user-adaptive interface, but they are just not used to online dictionaries incorporating those features. As a result, they have no basis on which to judge the usefulness of those features. (Müller-Spitzer, Koplenig & Töpel 2011, p. 270)

Customisability in the online version of the *Kollokationenwörterbuch* is therefore kept on a low level. Users should not be overwhelmed with settings to customise, or with too many features, but they should have certain possibilities to influence the behaviour of the user interface.

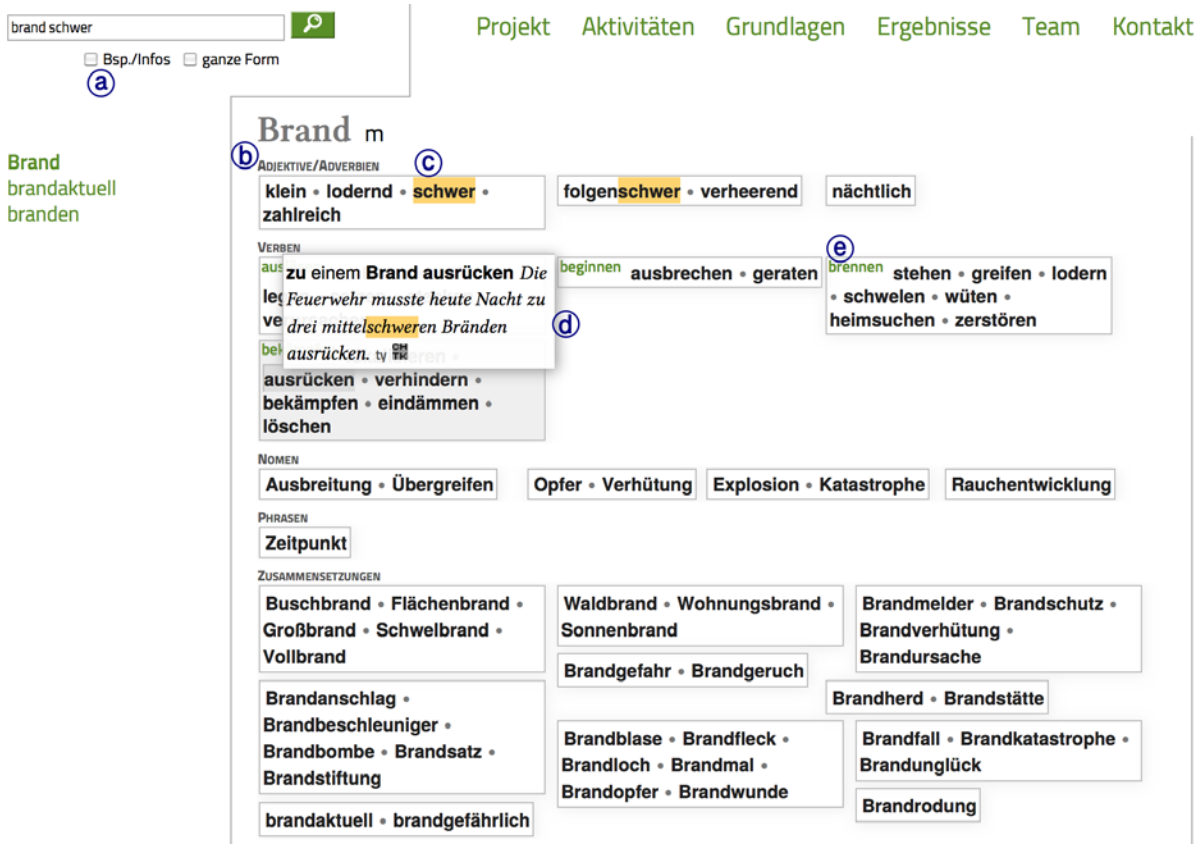

Figure 2: Dictionary article

Instead of the default two-level presentation outlined above, users can switch to a view where all information (examples, meaning indications, etc.) is displayed on one level (see Figure 2a). They can also toggle the display of collocates and extend citation forms. This gives users a view that resembles more the printed version of the dictionary.

4.6 Extensions

Possibilities to further extend the functionality of the dictionary[7] include, of course, the aforementioned linking of additional dictionary sources. Linkage from other (dictionary) sites to *Kollokationenwörterbuch* could also help to put it into a more general context, a bit detached from its status of a rather specialised dictionary and towards that of a tool commonly and readily used when writing. A Web service interface would greatly facilitate integration into other websites.

Another obvious enhancement, and probably the next step in further development, would be a version optimised for mobile devices, either as a mobile app on its own or just as a mobile-friendly version of the dictionary site.

Since a primary target audience of the *Kollokationenwörterbuch* are people producing text in writing, another promising possibility would be direct integration of the dictionary into text editors (as an add-on or plug-in). Just like they already get synonyms and spelling errors, authors could get collocates for a given word.

More ideas for extensions might come up with user feedback as soon as the dictionary site has been running for some time.

5. Conclusion

The *Kollokationenwörterbuch* is one of the first specialised collocations dictionaries for learners that has a dedicated online user interface. This user interface is the main topic of the present contribution.

Solutions have been proposed to two main problems of production-oriented collocations dictionaries. The problem of retrievability and navigation is tackled by a search facility over all the component words of the collocations, as well as with a two-level presentation that hides detail in the first step. Semantically motivated grouping is another feature likely to help in navigation.

The problem of information transfer, i.e. how to actually use a collocation that has been looked up, has been of great concern in the conception of the microstructure. Measures taken include explicit citation forms, meaning and usage indications and many example sentences.

[7] See also Roth (2012b).

In general, the online version of the *Kollokationenwörterbuch* should take the discussion on how collocations dictionaries should be presented online a step further.

6. References

Bickel, H., Gasser, M., Hofer, L. & Schön, C. (2009). Das Schweizer Textkorpus. In *Linguistik online* 39.3, pp. 5–31.

DiCE. Diccionario de colocaciones del Español. Accessed at: http://www.dicesp.com.

Duden online. Accessed at: http://www.duden.de.

DWDS. Digitales Wörterbuch der Deutschen Sprache. Accessed at: http://www.dwds.de.

elexiko. Online-Wörterbuch zur deutschen Gegenwartssprache. Accessed at: http://www.elexiko.de.

Fillmore, C. J., Kay, P. & O'Connor, M. C. (1988). Regularity and idiomaticity in grammatical constructions: The case of let alone. In *Language* 64.3, pp. 501–538.

Geyken, A. (2011). Die dynamische Verknüpfung von Kollokationen mit Korpusbelegen und deren Repräsentation im DWDS-Wörterbuch. In Klosa, A. & Müller-Spitzer, C. (eds.) *Datenmodellierung für Internetwörterbücher*. OPAL 2/2011. Mannheim: Institut für deutsche Sprache.

Häcki Buhofer, A. (2011). Lexikografie der Kollokationen zwischen Anforderungen der Theorie und der Praxis. In Engelberg, S., Holler, A. & Proost, K. (eds.) *Sprachliches Wissen zwischen Lexikon und Grammatik. Jahrbuch des Instituts für Deutsche Sprache 2010*. Berlin: De Gruyter, pp. 505–531.

Handl, S. (2009). Towards Collocational Webs for Presenting Collocations in Learners' Dictionaries. In Barfield, A. & Gyllstad, H. (eds.) *Researching Collocations in Another Language. Multiple Interpretations*. Basingstoke: Palgrave Macmillan, pp. 69–85.

Hausmann, F. J. (1985). Kollokationen im deutschen Wörterbuch. Ein Beitrag zur Theorie des lexikographischen Beispiels. In Bergenholtz, H. & Mugdan, J. (eds.) *Lexikographie und Grammatik. Akten des Essener Kolloquiums zur Grammatik im Wörterbuch, 28.–30.06.1984*. Tübingen: Niemeyer, pp. 118–129.

Herbst, T. (2009). Item-Specific Syntagmatic Relations in Dictionaries. In Nielsen, S. & Tarp, S. (eds.) *Lexicography in the 21st Century: In Honour of Henning Bergenholtz*. Terminology and Lexicography Research and Practice 12. Amsterdam & Philadelphia: John Benjamins, pp. 281–308.

Klein, W. (2004). Das digitale Wörterbuch der deutschen Sprache des 20. Jahrhunderts. In Scharnhorst, J. (ed.) *Sprachkultur und Lexikographie. Von der Forschung zur Nutzung von Wörterbüchern*. Frankfurt am Main: Lang.

Klosa, A., Schnörch, U. & Storjohann, P. (2006). ELEXIKO – A lexical and lexicological corpus-based hypertext information system at the Institut für Deutsche Sprache, Mannheim. In Marello, C. et al. (eds.) *Proceedings of the 12th EURALEX International Congress, Turin, Italy*, pp. 425–430.

Kollokationenwörterbuch. Typische und gebräuchliche Wortverbindungen des

Deutschen. Accessed at: http://www.kollokationenwoerterbuch.ch.

Le Fur, D. (ed.) (2007). *Dictionnaire des combinaisons de mots*. Paris: Le Robert.

LEO. Accessed at: http://www.leo.org.

Lo Cascio, V. (2012). *Dizionario combinatorio compatto Italiano*. Amsterdam: John Benjamins Publishing Company.

Makkai, A. (1972). *Idiom Structure in English*. The Hague: Mouton.

Müller-Spitzer, C., Koplenig, A. & Töpel, A. (2011). What Makes a Good Online Dictionary? – Empirical Insights from an Interdisciplinary Research Project. In Kosem, I. & Kosem, K. (eds.) *Electronic Lexicography in the 21st Century. New Applications for New Users. Proceedings of eLex 2011*. Ljubljana: Trojina, Institute for Applied Slovene Studies, pp. 203–208. Accessed at: http://www.trojina.si/elex2011/elex2011_proceedings.pdf.

OCDSE (2009). *Oxford Collocations Dictionary for Students of English*. Compiled by C. McIntosh. 2nd ed. Oxford: Oxford University Press.

Quasthoff, U. (ed.) (2011). *Wörterbuch der Kollokationen im Deutschen*. Berlin: De Gruyter.

Roth, T. (2012a). Using Web Corpora for the Recognition of Regional Variation in Standard German Collocations. In Kilgarriff, A. & Sharoff, S. (eds.) *Proceedings of the Seventh Web as Corpus Workshop (WAC7). Pre-WWW2012 Workshop, 17 April, 2012*, pp. 31–38. Accessed at: https://sigwac.org.uk/raw-attachment/wiki/WAC7/wac7-proc.pdf.

Roth, T. (2012b). *Wortverbindungen und Verbindungen von Wörtern. Lexikografische und distributionelle Aspekte kombinatorischer Begriffsbildung zwischen Syntax und Morphologie*. PhD thesis. Universität Basel.

Rundell, M. (ed.) (2010). *Macmillan Collocations Dictionary*. Oxford: Macmillan Education.

Siebenhüner, S. (2010). *Kollokationenwörterbuch: Schulstudie*. Universität Basel, Praktikumsbericht. Unpublished.

Sinclair, J. (1966). Beginning the Study of Lexis. In Bazell, C. E. et al. (eds.) *In Memory of J. R. Firth*. London: Longman, pp. 410–430.

Steyer, K. (2000). Usuelle Wortverbindungen des Deutschen. In *Deutsche Sprache* 2/00, pp. 101–125. *Wortschatz-Portal Universität Leipzig*. Accessed at: http://wortschatz.uni-leipzig.de/.