

## GIMME CONTEXT – TOWARDS NEW DOMAIN-SPECIFIC COLLOCATIONAL DICTIONARIES

Sylvana Krausse

Leiterin des Sprachenzentrums, Fachhochschule Nordhausen

E-mail: [krausse@fh-nordhausen.de](mailto:krausse@fh-nordhausen.de)

*The days of traditional drudgery-filled lexicography are long gone. Fortunately today, computers help in the enormous task of storing and analysing language in order to condense and store the found information in the form of dictionaries. In this paper, the way from a corpus to a small domain-specific collocational dictionary will be described and thus exemplified based on the example of the domain-specific language of mining reclamation, which can be duplicated for other specific languages too. So far, domain-specific dictionaries are mostly rare as their creation is very labour- and thus cost-effective and all too often they are just a collection of terms plus translation without any information on how to use them in speech. Particular small domains which do not involve a lot of users have been disregarded by lexicographers as there is also always the question of how well it sells afterwards. Following this, I will describe the creation of a small collocational dictionary on mining reclamation language which is based on the consequent use of corpus information. It is relatively quick to realize in the design phase and is thought to provide the sort of linguistic information engineering experts need when they communicate in English or read specialist texts in the specific domain.*

**Keywords:** collocational dictionary, corpus, domain-specific language, embedding, environmental engineering, general English, reclamation, storage of information.

**DOI:** 10.3846/1822-430X.2008.16.4.46–52

### Introduction and Object

The collocational and contextual information that is recorded in the dictionary stems from a corpus of 300,000 running words of various texts with topics which are related to remediation of abandoned mine sites. Originally, this small corpus was part of a bigger corpus (2 million words) of environmental engineering English which on top of mining reclamation also comprises soil and groundwater remediation and recycling technologies. Small subsections of the corpus thus serve to create a tailor-made collocational dictionary such as the predecessor of the described dictionary which is a dictionary on composting technologies featuring 107 head

entry words together with their collocational embedding and authentic sample sentences of the multi-word items the head entry is part of.

I do not want to go into detail on issues of representativity and range of corpora (cf. Williams 2002; Teubert 1999) but care has been taken regarding on the choice of texts. That is to say that corpus design is a crucial phase and texts that go into a corpus should be well selected to present as wide a range of text types and topics as possible. The mining reclamation subcorpus was then compared to a corpus of general English by the help of a corpus analysis software (in this case WordSmith 4) which allows users to derive a list of keywords which

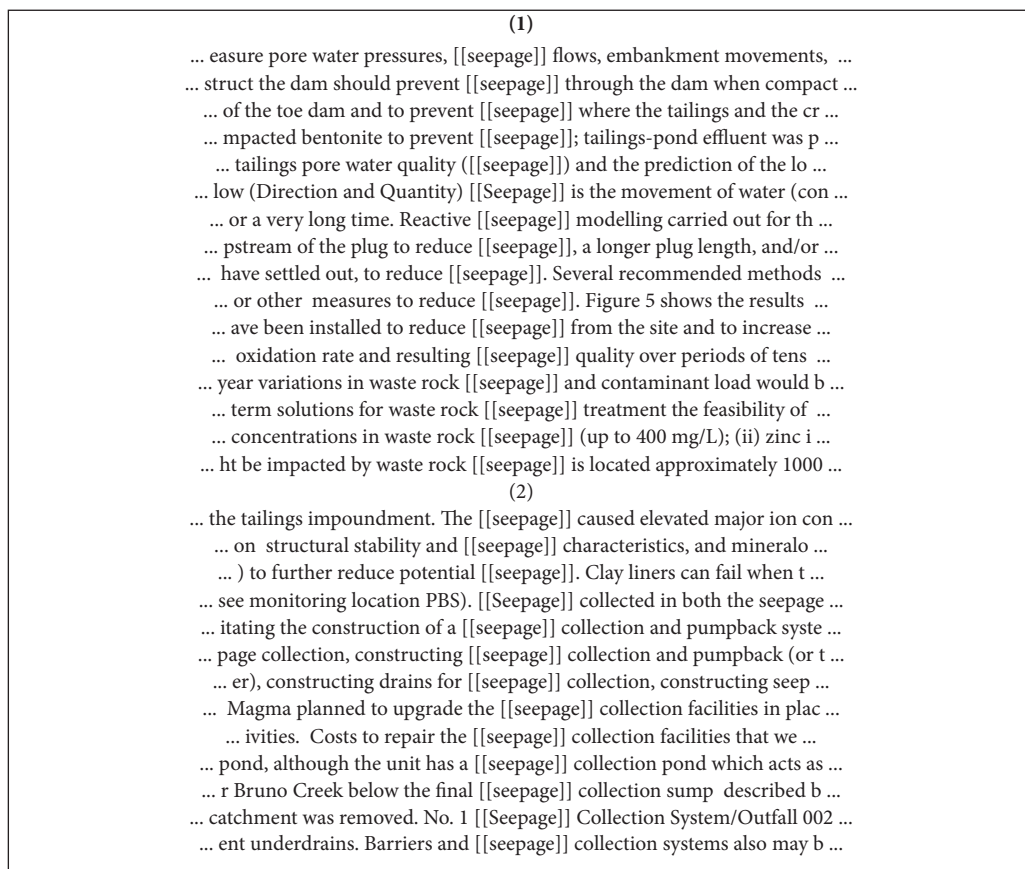


Fig.1. Concordance of *seepage*

are significantly more frequent in the specific language than in the general language. From this list, 286 words were chosen to serve as head entry words for the collocational dictionary on mining reclamation (the first ten words to appear in the dictionary are for instance: mine (noun and verb), mining, ore, waste, site, site-specific, pit, tailings, and impoundment). They are shown together with the collocational and context data which can be analysed with the help of another function of the above-mentioned software tool, the concordance function, which displays a search word or phrase in the following format:

All the lines in the corpus texts containing the search term, in this case *seepage*, are displayed in one window and the software allows

the user to sort alphabetically to the left (1) or to the right (2) of the search word. In this way, words which frequently occur together with the search term are made visible as they form patterns such as *prevent* or *reduce seepage* as well as *waste rock seepage* in (1) or *seepage collection facilities* or *systems* in the (2). It does not need to be stressed that the emphasis is on repeated events rather than on word combinations that occur only once. This is the kind of collocational information that finds entry in the dictionary as well as the full sample sentence to give the user the opportunity to see the functioning of the phrase in a normal sentence. How this is reflected in the dictionary entry will be shown in the next paragraph.

The Structure of a Dictionary Entry

Fig. 2 shows how the dictionary entry for the word *seepage* looks. It is number 54 of 286 head entries and is first listed together with a definition which situates the concept within the engineering domain. On the right there is the German translation of the head entry and underneath a sample sentence to illustrate how the word was used in one of the corpus texts. Below this sample sentence are all the multi-word units, *seepage* has been found to be part of and again on the right hand side the German equivalents of these multi-word units. In the last box underneath there are authentic sample sentences for each multi-word unit which can be seen printed in bold for better visualisation.

It can be seen that the dictionary entry format abstains from displaying all sorts of additional grammatical information in the form of *v.intr.*, *n.count.* etc., as well as any phonetic or etymological information, as even experienced dictionary users are said to often skip this information because looking up a term in a dictionary is often a very concise action where symbols and abbreviations describing dictionary categories are often ignored (cf. Bo-gaards 2003).

The Macrostructure

The ordering of the head entries in the dictionary has not been done alphabetically but according to its frequency in the corpus as well as to the meaning relationships between parti-

54. seepage n. def.: percolation of a liquid through soil or another medium	Sickerwasser, Versickern, Versickerung
Seepage was pumped from the deepest point of the pit and treated at a later stage of the process.	
1) contaminated seepage, 2) impoundment seepage, 3) minimize seepage, 4) control/discharge of seepage, 5) prevent seepage, 6) reduce seepage, 7) tailings seepage, 8) seepage collection, 9) seepage control, 10) seepage flow(s), 11) seepage from ..., 12) seepage pond(s), 13) seepage rate(s), 14) seepage through ..., 15) seepage water	1) kontaminiertes/verschmutztes Sickerwasser, 2) Sickerwasser/ Versickerung aus Aufstauung von Bergbau-abwässern, 3) Sickerwasser/ Versickerung minimieren, 4) Kontrolle/Austritt von Sickerwasser, 5) Versickerung verhindern, 6) Versickerung reduzieren, 7) Sickerwasser aus Abraum, 8) Auffangen des Sickerwassers, 9) Sickerwasser-kontrolle, 10) Sickerwasserfluss, 11) Sickerwasser aus, 12) Sickerwasserteich, 13) Sickerwasserrate, 14) Versickerung/Versickern durch ..., 15) Sickerwasser
The long-term potential of intrinsic remediation of <b>contaminated seepage</b> originating from a sulfidic tailings impoundment was assessed in the first phase of the project. ♦ Cycloning results in the deposition of the less permeable slimes behind the embankment, thus reducing <b>impoundment seepage</b> . ♦ Lining of impoundments can be a feasible option to <b>minimize seepage</b> and environmental impact. ♦ Synthetic liners are a relatively new development in the <b>control of seepage</b> in tailings impoundments. ♦ To ensure no <b>discharge of seepage</b> to the receiving waterbodies, a lined sump was installed further downstream. ♦ The materials used to construct the dam should <b>prevent seepage</b> through the dam when compacted. ♦ Diversion channels have been installed to <b>reduce seepage</b> from the site and to increase the amount of solution recovery. ♦ All <b>tailings seepage</b> eventually gets into the groundwater and discharges near the river. ♦ Avalanches damaged the liner during the construction making the construction of a <b>seepage collection</b> and pumpback system necessary. ♦ These threats to embankment stability may be partially reduced through <b>seepage control</b> . ♦ Records of <b>seepage flow</b> will indicate when significant changes occur. ♦ A seepage return pond was constructed to collect the <b>seepage from the tailings impoundment</b> . ♦ Current practices include the use of <b>seepage ponds</b> to collect seepage. ♦ As <b>seepage rates</b> accelerate, erosion of the dam material occurs thus leading to failure of the dam. ♦ Consolidation of the tailings in the impoundment is promoted by controlled <b>seepage through the dam</b> . ♦ The starter dam should be capable of passing <b>seepage water</b> .	

Fig.2. Dictionary entry of *seepage*

cular words. In this way, *seepage* is followed by *liner*, *lined* and *leakage* leading to other related words such as *solution*, *dissolve*, *solute* and *AMD*. If users look for a specific word, they have to first look it up in the register, where there are not just single words listed but also all the multi-word units the single words are part of. Moreover, the entries do not only comprise nouns as it is all too often done in terminological dictionaries but also verbs, adjectives and adverbs (see appendix). In this way, we can find the phrase *permeable layers* not just under the head entry *layer* but also *permeable* has got its own dictionary entry and consequently its own multi-word units of which it is part. The reasons for abandoning the admittedly more comfortable and more traditional method of alphabetical ordering will become clear in the following paragraph.

### Dictionary User and Dictionary Usage

The Collocational Dictionary of Mining Reclamation is intended for German-speaking engineering experts working in the field of restoring and remediating old mine sites. That is for people who use English as a lingua franca in the workplace or need to read specialist literature on mining reclamation; thus fairly advanced language learners. They dispose of a profound expertise in their field of engineering which helps them to place the words, the multi-word units and their sample sentences in their situational and conceptual context. What the dictionary helps them to understand is first of all to find the translational equivalent in either English or German (via the register function) but what is more important to show them how the words are used in English language, i.e. which words frequently occur together with the head entry, which verbs and prepositions go with it and how it has been used by proficient language users and field specialists in actual texts.

If it had been only for this loop-up function, an alphabetical order would have been the better choice but on top of looking up single words and phrases, the dictionary serves a di-

dactic purpose. It can be also used for language learning purposes. The entries are ordered according to meaning relationships. So users can start from one central entry of a topic they want to focus on and follow the other entries from there. Moreover, the sample sentences feature already many of the other related words which can be looked up under another entry and in this way, associations between concepts are created in the users' minds very much of the kind they are presented in the knowledge of this engineering branch in their native language. This is furthermore very much the way corpus work has been used for instance in courses for students of translation in the form of serendipity learning (cf. Bernardini 2000), which means that you start with a certain search term and its concordance, analyse the found language patterns and deduce rules about language use from the actual evidence in the corpus. One search will lead to another, and the new dictionary format allows this serendipity learning too. Here the serendipity is a guided one because the lexicographer has used the corpus evidence already to unveil meaning relations between the words which are thought to be pivotal for a certain language domain and there is thus a certain logic behind the ordering.

### The Added Value of Small Domain-Specific Collocational Dictionaries

If you are active in the field of language teaching, you will have probably encountered the following situation in one way or the other: a student looks up a word in a dictionary, and from all the meanings given chooses one that does not fit the context. The teacher corrects this mistake with the comment that in this very context, another word needs to be used. One of my students tried to justify himself after my correction lately, saying that if he had had a sample sentence this mistake would not have occurred... And how right he was! Usually general language dictionaries and even some more comprehensively orientated technical dictionaries in a single entry unite information

waste in the domain of <b>composting technologies</b>	shared concepts ( <b>in composting technologies and mining reclamation</b> )	waste in the domain of <b>mining reclamation</b>
biowaste	waste disposal	mine waste/ mining waste
greenwaste	waste generation	waste dump(s)
household waste	waste material(s)	waste pile(s)
industrial waste		waste rock (dump(s))
landscape waste		waste rock pile(s)
municipal waste		waste stream(s)
(enter) the waste stream		
organic waste (materials)		
prevent waste		
receive waste, etc.		

Fig. 3. *Different collocational contexts of waste*

on different domain labels of a word and usage information. It is then up to the skillfulness of the dictionary user to locate the wanted information. The described type of dictionary is geared to be used just in one domain only and removes the task of understanding and applying domain labels correctly from the dictionary user (moreover, domain labelling is often not done to the required degree). In Fig. 3 you can see which typical multi-word units *waste* is part of in the domain of composting technologies (1<sup>st</sup> column) and which it is part of in the domain of mining reclamation (3<sup>rd</sup> column). It is only the collocations in the 2<sup>nd</sup> column that occur in both domains.

This shows that the linguistic co-text of one word (*waste* in this case) differs widely from one domain to the other. In the hands of an engineering specialist in this domain, this can be a decisive advantage.

Conclusion

In this paper I aimed to illustrate the underlying principles and the advantages of a domain-specific dictionary based on the example of a dictionary on mining reclamation phrases. I have avoided using the word ‘terminology’ in the previous sentence as it is not just terms which are recorded in this dictionary but all kinds of collocational and colligational information which was derived from a corpus analysed with the help of a concordance tool. This kind

of dictionary constrains itself to displaying just phraseological information on domain-specific frequent English words relating to topics dealt with in texts on mining reclamation and their German translational equivalents. One crucial element of each dictionary entry are the authentic sample sentences derived from corpus work.

Such a collocational dictionary can serve two functions: first, the normal look-up function of phraseological and translational information, and second, it may serve as language learning material on a higher level to that which can be gained from normal dictionaries. Users can start at one central entry of the topic they are currently dealing with and go on from there to meaning related entries which possess a statistical likelihood to occur in the context of the topic. For instance, in texts about *backfilling* of mines, it is likely one will encounter the following words: *layers, compaction, compacted, hydraulic, consolidation, consolidated* and of course the whole word family of *backfilling* such as *backfill* n., *backfill* v., *backfilled*, etc. In this way, associative thinking and language learning is supported very much by the way that learning in general happens, that is to say in the situational context.

The dictionary is bilingual and unidirectional: it may serve the German engineering specialist coping with understanding English specialist texts and finding phraseological

information needed to produce his or her own specialist language. At the same time, the dictionary format lends itself to being easily adaptable to other language pairs: the English frame/foundations may stay the same and just the head entry and its multi-word units are translated to other languages. This as well as its computer-based creation process, which allows one to work relatively quickly on language domains which evolve and develop relatively fast, might turn out to be a crucial benefit of this dictionary type.

## References

- Bernardini, S. 2000. *Competence, Capacity, Corpora. A Study in Corpus-Aided Language Learning*. Bologna: CLUEB.
- Bogaards, P. 2003. "Uses and Users of Dictionaries", in P. van Sterkenburg (Ed.). *A Practical Guide to Lexicography*. Amsterdam/Philadelphia: Benjamins, 26–33.
- Krausse, S. (in print). *Collocational Dictionary of Mining Reclamation*. Nordhäuser Hochschultexte. Nordhausen.
- Krausse, S. (in print). *Collocational Dictionary of Composting Technology*. Nordhäuser Hochschultexte. Nordhausen.
- Teubert, W. 1999. *Corpus Linguistics-A Partisan View*. TELRI Newsletter 8/99, 4–19.
- Williams, G. 2002. In Search for Representativity in specialised Corpora – Categorisation through Collocation, *International Journal of Corpus Linguistics* 7(1): 43–64.

## Appendix

dictionary entry for *permeable*

85. <b>permeable</b> adj. def.: allowing liquids to pass through	<b>durchlässig, permeabel</b>
Fine tailings are intercalated with waste rock, which is a highly porous and <b>permeable</b> medium.	
1) highly/less permeable, 2) permeable layer(s), 3) permeable sand(s)	1) sehr/weniger durchlässig, 2) durchlässige Schichten, 3) durchlässiger Sand
Most leach sites take advantage of existing, <b>less permeable</b> surfaces and utilize the natural slope valleys for the collection of pregnant leach solution. ♦ The transition zone is typically consolidated due to significant lateral drainage through <b>permeable layers</b> . ♦ Most of the precipitation at the site infiltrates through <b>permeable sands</b> and gravels.	

dictionary entry for *leach*

59. <b>leach</b> (out) v. def.: to remove soluble particles with the help of an acidic solution	<b>auslaugen, auswaschen</b>
Contaminants in mine waste can leach out into surface and groundwater causing serious pollution that can last for many years. ♦ Sulfuric acid may leach metals, yielding a metal-rich, acidic solution.	

dictionary entry for *downstream*

244. <b>downstream</b> adj./adv. def.: towards where a liquid flows	<b>flussabwärts</b>
Overflow is initially contained by the dam at pond, located about 500 feet <b>downstream</b> from the raffinate pond. ♦ This work predicted a seepage of 11 l/s into the tunnel <b>downstream</b> of the plug with 100 m of head.	



dictionary entry for AMD

70. AMD or ARD Acid Mine Drainage/ Acid Rock Drainage def.: acidic water escaping from abandoned mines	Saures Grubenwasser, Lettenwasser
When <b>AMD</b> occurs, treatment may be required for centuries.	
1) control AMD/ARD, 2) mitigate AMD/ARD, 3) potential for AMD/ARD, 4) predict AMD/ARD, 5) prevent AMD, 6) AMD/ARD generation, 7) AMD potential, 8) AMD/ARD production, 9) ARD source	1) saures Grubenwasser überwachen/kontrollieren, 2) saures Grubenwasser vermindern, 3) Gefahr der Bildung von saurem Grubenwasser, 4) saures Grubenwasser vorherberechnen, 5) saures Grubenwasser vermeiden, 6) und 8) Bildung von saurem Grubenwasser, 7) Potenzial/Gefahr der Bildung von saurem Grubenwasser, 9) Herkunft/ Herkunftsort) des sauren Grubenwassers
Water diversion, soil covers and plastic liners, dewatering, inundation, underground mine sealing, barriers are water management techniques used for <b>controlling AMD</b> . ♦ Predicting the <b>potential for AMD</b> is currently difficult, expensive, and of questionable reliability. ♦ The addition of basic industrial residues to layered co-mingling can <b>mitigate AMD</b> . ♦ New regulations require testing of rock using a procedure for <b>predicting AMD</b> . ♦ The most reliable strategy for <b>preventing AMD</b> is to submerge the waste rock or tailings under water to prevent exposure to air. ♦ Tailings will be mounded so that future precipitation runs off the sides and thereby reduce the potential for <b>AMD generation</b> . ♦ Static testing is the first step in understanding <b>AMD potential</b> . ♦ When the process of <b>AMD production</b> has started, it is very difficult, even impossible to stop. ♦ Solutions to ARD-caused problems usually focus on preventing the contact of water with the <b>ARD</b> source.	

NAUJŲ KONKREČIOS SRITIES ŽODŽIŲ JUNGINIŲ ŽODYNŲ LINK

Sylvana Krausse

Tradicinio nuobodaus leksikografinio darbo laikai yra praėję. Laimei, šiandien kompiuteriai atlieka tą milžinišką darbą, kaupdami ir nagrinėdami kalbą, kad kauptų ir teiktų informaciją, randamą žodynuose. Šiame straipsnyje yra aptariamas kelias nuo pagrindinio žodyninio fondo iki konkrečios srities žodžių junginių žodyno, pasiremiant konkrečios srities pavyzdžiais. Tą patį galima taikyti ir daugeliui kitų sričių. Ligi šiol konkrečios srities žodynų nėra daug, nes tokio žodyno sudarymas reikalauja daug darbo, yra brangus ir dažniausiai tėra tik terminų rinkinys, su vertimu, be jokios informacijos, kaip juos vartoti kalboje. Leksikografai dažnai nekreipia dėmesio į ypač neesmines sritis, nes visuomet iškyla klausimas, ar žodyną pirks, kai jis bus išleistas. Čia bus aptariama, kaip sudaryti žodžių junginių žodynėlį pasitelkiant pagrindinio žodyninio fondo informaciją. Šį pirmąjį darbo tarpinį, palyginti, yra lengva įvykdyti. Manoma, kad toks žodynas technikos žinovams, inžinieriams teiks lingvistinę informaciją, kurios prisireikia, kai jie bendrauja angliškai arba kai skaito konkrečios specialybės tekstus.

**Reikšminiai žodžiai:** žodžių junginių žodynas, pagrindinis fondas, konkrečios srities kalba, įtvirtinimas, aplinkosauga, bendrinė anglų kalba, perėmimas, informacijos kaupimas.

Įteikta 2008-02-11; priimta 2008-02-15