International Journal of Lexicography, Vol. 28 No. 3, pp. 299–317
doi:10.1093/ijl/ecv019

LEXICOGRAMMATICAL PATTERNS AND CORPUS EVIDENCE IN SCHEMANN'S DICTIONARY OF IDIOMS

Alexander Geyken: *Berlin-Brandenburgische Akademie der Wissenschaften (geyken@bbaw.de)*

Abstract

This article presents a strategy for mapping entries in Schemann's dictionary of German set phrases to lexicogrammatical patterns. With more than 10,000 headwords and 33,000 entries, Schemann's dictionary is the most comprehensive printed dictionary of German set phrases. The result of the mapping yields more than 8,500 distinct pattern classes, which are used to generate effective corpus queries. A qualitative evaluation of these queries with respect to the 4 billion word corpus of the German DWDS project shows how the entries can be improved with corpus evidence and how the dictionary can be augmented with corpus examples.

1. Background

Lexicogrammatical patterning, characterized as the classification of words not only on the basis of their meaning but also on the basis of their co-occurrences with other words, has a long tradition for the investigation of the lexis (e.g. Firth 1957, Halliday 1966). In this tradition co-occurrences are regarded as significant if they co-occur more often than dictated by chance (Sinclair 1966). Some 20 years later, with the advent of second generation corpora, John Sinclair was the first to implement these ideas with the corpus-driven Cobuild project. Sinclair (1991) also formulated the idiom principle, according to which much of what we say and write is based on semi-preconstructed phrases with grammar being used as a fall-back resource when we run out of suitable semi-preconstructed phrases. Phraseologists and corpus linguists working on phraseology agree in assuming a continuum between completely fixed expressions (e.g. *kick the bucket*) and full grammatical variation (*give sth. to sb.*) where the role fillers are open lexical (or phrasal) sets. However, where

phraseologists can base their classification on semantic opacity and on frozenness of form (Frazer 1970, Burger 2007), corpus-linguistics can only relate form-based properties to frequency. Indeed, a major difficulty for corpus-based approaches to phraseology is that they fail to establish typologies of patterns that distinguish free grammatical constructions from 'phraseological' patterns (Moon 2007). In this article we describe phraseological patterns from a different perspective, taking a large dictionary of fixed expressions as a point of departure. The hypothesis is that a comprehensive collection of fixed expressions in a language can shed a different light on the way in which fixed expressions and lexicogrammatical patterns are interrelated and how they are distributed according to frequency of occurrence.

One dictionary that meets these conditions is the comprehensive dictionary of German set phrases, Schemann's *Deutsche Idiomatik* (Schemann 1993, 2011). With more than 33,000 different entries and 10,000 headwords, Schemann's *Deutsche Idiomatik* is currently by far the most comprehensive dictionary of set phrases for German. It exceeds by more than three times the renowned dictionary of set phrases published by Duden (Duden 1992, 2013), which comprises around 10,000 entries. However, in contrast to Duden where each entry has a semantic paraphrase and one or more corpus examples, Schemann's dictionary has a less ambitious information programme. His main goal is to record as many different phraseological expressions as possible in order to support language learning (Schemann 2011). Schemann's dictionary has been influential for bilingual lexicography: Bilingual versions on the basis of his dictionary were compiled for English as well as for French, Italian, Portuguese, and Spanish (Mellado Blanco 2014). For all these dictionaries the *Deutsche Idiomatik* works as the pivot language, i.e. all bilingual dictionaries provide translations into the respective languages for each entry. Thus these bilingual dictionaries can be used for comparative studies of idiomatic expressions in these languages. An evaluation of Schemann's dictionary, including the up-to-dateness, the completeness as well as corpus evidence of the entries in large corpora would therefore be useful for both the German monolingual dictionary and its bilingual derivatives.

Terminology in this article is used as follows. *Lexicogrammatical patterns* are defined as sequences of pairs of lexemes and part-of-speech-labels together with additional phrasal information and morpho-syntactic constraints, and we will use the neutral term *set phrase* to denote collocations, idioms, support verb constructions, quasi-idioms, catch-phrases, routines and proverbs.

The present article is organized as follows. We briefly describe the way the entries are organized in Schemann's dictionary (section 2). In section 3 we present our strategy for mapping entries in Schemann's dictionary to lexicogrammatical patterns. We also provide some results of this work, including a distribution of the lexicogrammatical patterns according to frequency of occurrence (section 4). In section 5 we show how these patterns can be converted

into corpus queries and how this process influences the precision and recall in corpora. The conclusion in section 6 will summarize the results and give an outlook on the next step of our work in which we plan to convert all the lexicogrammatical patterns compiled in section 3 into corpus queries semi-automatically.

2. Schemann's dictionary of German Idiomatics

Schemann published his dictionary of *Deutsche Idiomatik* first in 1993 (with Klett) and a revised version in 2011 (with deGruyter) together with an electronic pdf-version. It is worth noting that the revision is limited to the introduction, the content remains unchanged. Schemann uses the notion of *Idiomatik* in a very broad sense. The selections range from completely frozen expressions to multi-word expressions where all components are substitutable. The complexity of form varies from two-word expressions to full sentential clauses. The dictionary thus contains collocations, support verb constructions, proverbs as well as semantically opaque idioms and formulaic expressions. The entries are arranged alphabetically under a headword. Here the language intuition of the average dictionary user is reflected rather than a linguistic classification. Particular emphasis is put on the larger context of the idiom. A context can be either linguistic (morpho-syntactic restrictions, syntagmatic variations) or related to the pragmatic use of an entry. (1) provides a snippet from the dictionary for the headword *Gehör* (hearing) with the first entry subsumed under the headword, *ein absolutes Gehör*/ *das absolute Gehör haben*. The entry is followed by example sentences illustrating function and meaning of the entry.

(1) GEHÖR: **ein absolutes / das absolute Gehör haben**. Wer das absolute Gehör hat … –, muß der auch besser interpretieren? – Nein.
literally: an absolute / the absolute hearing have
english: to have perfect pitch
example: sb. with a perfect pitch … should he/she be able to interpret better? No.

Entries in the dictionary are more than a mere linear sequence of tokens. Rather, they correspond to a set of syntagmatic contexts around the target word. Examples (1-4) demonstrate the lexicographic shorthand used by Schemann: **/** to indicate a mutually exclusive choice between elements to the left and right of the symbol (1, 3-4), **()** to indicate an optional phrase that may be omitted, dots … to designate arbitrary clauses, and placeholders such as *jdn.* (s.o.) or *etw.* (sth.*)* to designate flexible components that allow substitution, in this case an accusative noun phrase (4). The dictionary uses diasystematic markers (e.g. [coll.] in (3,4)) to specify if the entry is colloquial, pejorative,

rare or obsolete, or used only in a particular domain. The bold typeface is employed to mark the mandatory part of the entry. In the context of a slash the part in bold is preferable to its non-marked alternative (e.g. in (3) ***sich kein*** vs. *nicht ein*).

(2) **einen** **zischen**
one fizzle
to knock back a drink, to have a swift one

(3) **sich kein**/nicht ein **X für ein U vormachen lassen** [coll.]
Self no / not one X for a U show let
lead someone down/up the garden path

(4) (voll) **auf** jdn./... **abfahren** [coll.]
(fully) on sb. / ... go off
to really fancy somebody

The basis for the entries is a reading programme of selected literature (about 50 works of well-known German authors, including Böll, Hesse, Th. Mann, Musil, and Walser), of regular reading of daily and weekly newspapers and magazines (including *Die Frankfurter Allgemeine Zeitung*, *Die Süddeutsche Zeitung*, *Die Zeit*, *Merian*, *Bravo* and *Der Spiegel*), of 'personal observations of spoken language', and finally of large documentary reference dictionaries of contemporary German (*Duden, Wahrig*). The example sentences provided in the dictionary are examples made-up by the author. Their aim is to provide an authentic semantic, pragmatic and stylistic context for the dictionary entry. Schemann notes in his introduction (Schemann 1993, 2011) that he would have preferred to provide corpus examples, but acknowledges the fact that providing appropriate corpus examples for more than 30,000 entries was beyond his capabilities as a lexicographer. Furthermore he states that all other works (in German) providing corpus examples for idioms were all significantly smaller than his dictionary.

3. Transforming entries in Schemann's dictionary into lexicogrammatical patterns

The entry structure of Schemann's dictionary is not immediately usable for a transformation into lexicogrammatical patterns. A number of processing steps were carried out to transform the entry structure into a more suitable machine-readable format. The first step consists of pre-processing: Entries with optional elements are rendered simultaneously as multiple patterns, a pattern with the optional element and a pattern without. Likewise, the symbol [...] can be thought of as an unconstrained substitution. Hence, it is possible to remove the token altogether without loss of significant information. A more difficult problem is the resolution of the scope of the **/** operator. It is not specified

explicitly in (2) whether the scope of / should be [*ein absolutes*]/ [*das absolute*] or *ein* [*absolutes*/*das*] *absolute Gehör haben*. Generally, even though this can be done very quickly, such disambiguation requires manual decision on a case-by-case basis.

As a result of step one each entry is transformed into one or more elementary entries, i.e. entries without structural ambiguities (cf. section 4). The second step consists of annotating these patterns with part-of-speech tags. For example, the word *einen* in (1) may be interpreted as a determiner, a verb, or a cardinal number, while *das* in (2) is either a determiner or a relative pronoun. Given the 33,000 entries and an average length of 4.3 tokens for each entry, manual association of each token to its part-of-speech tag (POS-tag) would not be feasible. In Geyken and Boyd-Graber (2003) we describe a semi-automatic method for labelling the entries of Schemann's dictionary with POS-tags. In the present article, we summarize the main steps. We used a machine learning approach that involved training on the specific entry structure of the dictionary because parsing the entries of idiom dictionaries is not equivalent to the parsing of naturally occurring text: the entries follow a formalized structure quite different from that of ordinary language, the tokens are generally lemmatized and the verb is placed at the end of the entry. As a training set for the machine learning approach, we annotated a subset of the idiom dictionary of 6,000 entries with POS-tags. The set of POS-labels used for this task was a simple set of 10 tags: *Adj, Adv, Conj, Det, N, NA, Prep, Pron, Ptk, V,* [NOTE 1]. In addition a list of plausible patterns (Tschichold 2000) was elaborated. The goal was twofold: each entry of the idiom dictionary should be associated with one of those pre-established classes of POS-sequences and each token should be annotated with its POS-tags. On the basis of the above-mentioned training set the result (the training set equals the test set) was as follows: the tagger yielded an accuracy of 97.5% for the labelling with POS-tags, and 86% recall of associations to the POS-sequences.

However, this approach had two limitations: the recall of 86% is far from being satisfactory and the syntactically reduced tag-set results in some loss of quality. An example of the drawbacks of applying the small tag-set is the following dictionary entry where the distinction between indefinite pronouns and reflexive pronouns is useful.

(5) sich ein paar Tränen abquetschen
itself a few tears squeeze
to squeeze out a few tears

In (5) the correct mapping of the tokens *sich* and *paar* should be *sich* (reflexive pronoun) and *paar* (indefinite pronoun). With just the reduced tag-set at our disposal we would have to label both *sich* and *paar* with PRON, which again would be syntactically mapped to the meaningless pattern' pron.det.pron.N.V'.

Because of these limitations we decided to use STTS, a tag-set widely used for annotating German text corpora with part-of-speech labels (Schiller et al. 1995). The STTS tag-set consists of 54 grammatical categories. In this article the following tags will be used: ADJA (attributive adjective), ADV (adverb), APPR (attributive pronoun), ITJ (interjection), KON (coordinating conjunction), PAV (pronominal adverb), PDS (substituting demonstrative pronoun), PIDAT (indefinite pronoun with determiner in attributive position), PIS (substituting indefinite pronoun), PPOSAT (attributive personal pronoun), PRF (reflexive personal pronoun), NE (proper noun), NN (normal noun), VAINF (auxiliary verb, infinite form), VMINF (modal verb infinite), VVINF (full verb infinite), VVFIN (full verb finite), and VVPP (full verb past participle). Furthermore we split up the STTS tag ART (article) into ARTDEF and ARTINDEF (definite and indefinite article, respectively), and we use a generic class PTK (particle) for all occurrences of any particle.

The recategorization of these POS-sequences on the basis of the STTS tag-set is the third step of our work. It was carried out as a correction task of the pre-classification step 2 described above [NOTE 2]. In addition we annotated morpho-syntactic constraints which are imposed by the context, such as a particular value for the number feature of the noun or a specific inflected form of the verb. We also performed a markup of noun phrases $[]_{NP}$, adjective phrases$[]_{AP}$, prepositional phrases$[]_{PP}$ and clauses $[]_{CL}$ if they denote adjacent lexeme sequences in the entry. Furthermore we marked up morpho-syntactic constraints including restrictions on number and person. In (6) to (16) we use the shorthand $ to express that a lexeme can occur in all its morphological variations. If the lexeme is unmarked it can only be used in the literal form. The motivation for this recategorization was twofold. A correct classification of POS-patterns according to the STTS tag-set would provide better insights into the number and distribution of these patterns in idiom dictionaries (cf. section 4), and, secondly, the phrase tags are important for the transformation of patterns into corpus queries. Some examples illustrate this third step.

(6) **sich steif und fest einbilden**, daß [...]
self stiff and fast imagine, that [...]
to have it stuck into one's head that...
(6.1) sich|REFLPRON [steif|ADV und|KON fest|ADV]$_{ConjP}$ $einbilden|V

(7) mit jdm. **Tacheles reden**
with sb. Tacheles(yidd.) speak
To talk straight with somebody
(7.1) mit|APPR [NA] Tacheles|NN $reden|VVFIN

(8) **von A bis Z** Unsinn / erlogen / erfunden / [...] sein
from A to Z nonsense / lied / invented be
sth./it is a pack of lies from A to Z

(8.1) [von|APPR A|NN bis|APPR Z|NN]$_{PP}$ Unsinn|NN $sein|VAINF
(8.2) [von|APPR A|NN bis|APPR Z|NN]$_{PP}$ erlogen|VPP $sein|VAINF
(8.3) [von|APPR A|NN bis|APPR Z|NN]$_{PP}$ erfunden|VPP $sein|VAINF

In (6) and (8) subordinate clauses, placeholders for noun phrases, and the unconstrained substitution [...] were deleted. Only when the placeholders were part of a prepositional phrase were they preserved and replaced by the tag [NA], as in (7). In (8), the '/' operator stands for mutually exclusive choices. Therefore the initial entry was split up into three patterns (8.1) – (8.3).

(9) **das**/was j. sagt/... **ist** (ja/doch) (alles) **kalter Kaffee**
That/what sb. says/... is (yes/yet) (all) cold coffee
That's an old hat!
(9.1) etw. (ja/doch) (alles) [**kalter Kaffee**]$_{NP}$ sein
(9.2) (das|PDS) (ja|PTK) (alles|PIS) [kalter|ADJA Kaffee|NN]$_{NP}$ ($sein|VAINF)
(9.3) (das|PDS) (doch|PTK) (alles|PIS) [kalter|ADJA Kaffee|NN]$_{NP}$ ($sein|VAINF)

(9) is a case where the original dictionary entry was considered inappropriate for further processing. The dictionary claims that the tokens *das ist* and *kalter Kaffee* are mandatory elements. However, as can be easily verified with corpus examples, neither the demonstrative pronoun *das* (that) nor the auxiliary *sein* (to be) are mandatory. (9.2 and 9.3) shows the result of this transformation. The nucleus of (9) is the adjacent NP *kalter Kaffee* where the adjective and the noun cannot be modified.

(10) is an example for an entry where the exact string of the mandatory element has to be matched. There is no inflection or different word order possible. The entry contains the string exactly as it has to be used with the verb in 3rd person indicative and the noun in the singular form.

(10) (**das**) **weiß der Henker** [ITJ]
(that) knows the hangman
who knows
(10.1) weiß|VVFIN [der|ARTDEF Henker|NN]$_{NP}$
(10.2) (das|PDS) weiß|VVFIN [der|ARTDEF Henker|NN]$_{NP}$

4. The wealth of patterns in Schemann's dictionary

Schemann's dictionary consists of 10,392 different headwords that give rise to a total number of 33,237 entries. Figure 1 shows that the frequencies of headwords and entries follow a Zipf-like distribution. Very few headwords have a large number of entries. The top 3 are *machen* (to make), *Kopf* (head) and *Hand* (hand) with 291, 250 and 234 entries respectively. The list of the top 25

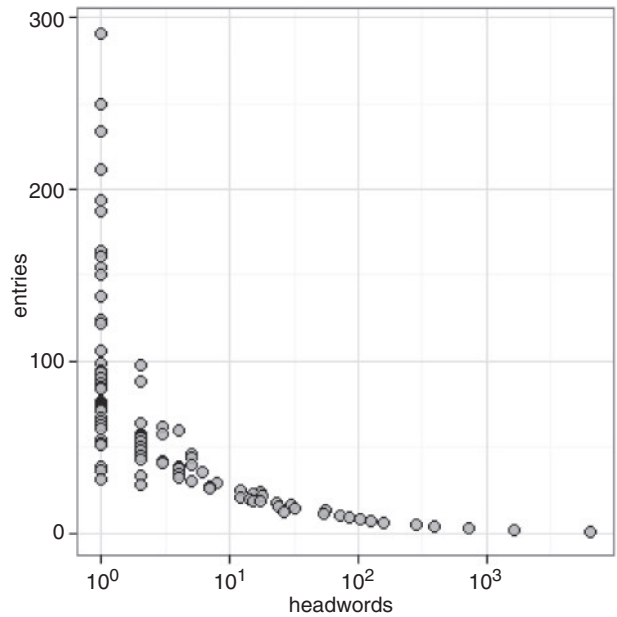

Figure 1: Distribution of headwords and entries by frequency; the x-axis is plotted in a logarithmic-scale.

headwords is displayed in appendix 1. It is not surprising that among the top 25 are high-frequency light verbs like *make* and *become* and nouns that relate to body parts (*head*, *foot*) as well as abstract nouns (*life*, *death*). In fact only 20 headwords have more than 100 entries. The vast majority, i.e. 6,329 (resp. 1,625) headwords just stand for one (resp. two) entries.

The 33,237 dictionary entries were transformed into 43,337 lexicogrammatical patterns by applying the method described in section 3. As stated above, the number of lexicogrammatical patterns is higher than the number of entries since an entry may be decomposed into several elementary entries which in turn have a unique mapping to its corresponding lexicogrammatical patterns.

Figure 2 displays the Zipf-like distribution between the total number of lexicogrammatical patterns (43,337) and the number of distinct pattern classes (8,775). We use the term *pattern class* (or short *class*) to refer to the projection of a lexicogrammatical pattern to its POS-sequence. For example, the lexicogrammatical pattern *take*|VV *a*|ARTINDEF *decision*|NN is mapped to the pattern class VV ARTINDEF NN. The plot shows that the number of lexicogrammatical patterns per pattern class does not drop as quickly as it does for the number of entries per headword in Figure 1. Appendix 2 lists the top 25 pattern classes. The most frequent class is NN V (freq = 659). This class subsumes light verb constructions with no determiner such as *Amok laufen* (run amok) or *Abhilfe schaffen* (find a remedy). It is interesting to note that this type outperforms both the classes with a definite determiner (rank 2, freq = 621)

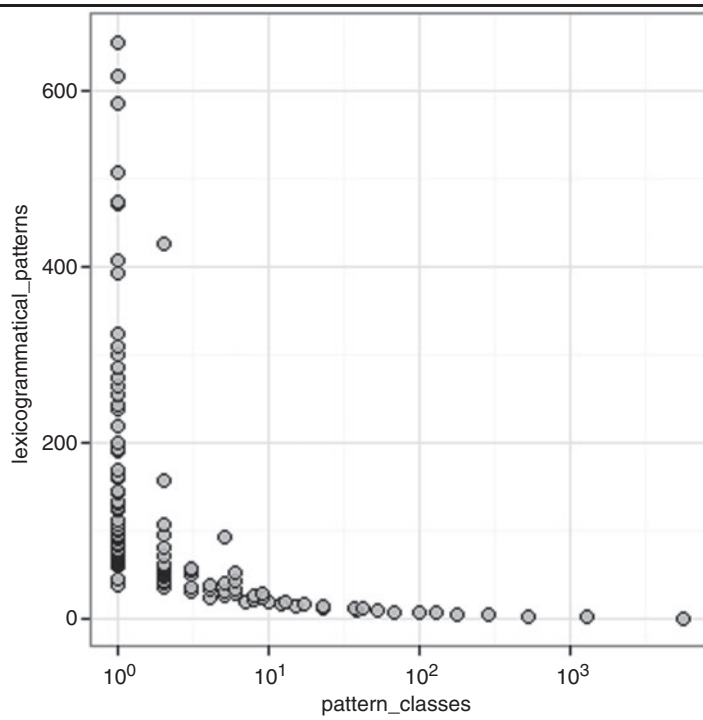

Figure 2: Pattern classes and lexicogrammatical patterns by frequency; x-axis is plotted in a logarithmic-scale.

and that with an indefinite determiner (rank 8, freq = 427). All classes among the top ten are verb-noun classes. Only one class among the top ten contains a modal verb. The first phrasal pattern classes are ARTDEF ADJA NN, ranked 18th with a frequency of 243, followed closely by the class starting with a preposition, notably APPR ADJA NN (rank 20, freq = 220). Examples for these phrasal classes are *das ewige Eis* (eternal ice) and *in gereiftem Alter* (at a mature age). At the end of the ranked list are rare classes. There are 5,661 unique pattern classes and 1,303 classes with only two lexicogrammatical patterns. Examples for pattern classes that occur only once are [PINEG NN]$_{NP}$ [APPR NA]$_{PP}$ [APPRART NN]$_{PP}$ VVINF VMINF and [PINEG NN]$_{PP}$ PAV [APPRART NN]$_{PP}$ VVINF VMINF. These pattern classes correspond to the lexicogrammatical patterns (11.3, 11.4) that result from the decomposition of (11) into elementary entries (11.1, 11.2).

(11) **mit** e-r S./damit **kannst du**/kann er/. . . **keinen Hund hinterm/hinter dem Ofen hervorlocken**/. . . lockt man keinen Hund . . . hervor
no dog with sth./therewith behind the oven lure out can
This is nothing to write home about

(11.1) keinen Hund mit etw. hinter dem Ofen hervorlocken können

(11.2) keinen Hund damit hinter dem Ofen hervorlocken können

(11.3) [keinen|PINEG Hund|NN]$_{PP}$ [mit|APPR NA]$_{PP}$ [hinter|APPR dem|ARTDEF Ofen|NN]$_{PP}$ hervorlocken|VVFIN $können|VMFIN

(11.4) [keinen|PINEG Hund]$_{NP}$ damit|PAV [hinter|APPR dem|ARTDEF Ofen|NN]$_{PP}$ hervorlocken|VVFIN $können|VMFIN

5. Matching patterns with corpus examples

In section 3 we described how we (manually) transformed the dictionary entries of Schemann's *Deutsche Idiomatik* into machine-readable lexicogrammatical patterns. In this section we show how these patterns can be converted into corpus queries. These corpus queries generally do not provide exact matches with corpus sentences since set phrases may reveal more variation than generally expected (Fellbaum 2007). However, by taking advantage of the information provided by the annotated patterns, it is possible to reduce the number of false positives while preserving a reasonable recall.

The underlying text corpus for this analysis is the corpus of the DWDS project (www.dwds.de) of the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW). The DWDS project comprises four subcorpora: the DWDS-Kernkorpus of the 20th/21st century, a balanced reference corpus of 110 million tokens (Geyken 2007), a balanced historical corpus currently comprising 120 million tokens for the period from 1600 to 1900, compiled at the BBAW for the project *Deutsches Textarchiv* (DTA, German Text Archive, www.deutschestextarchiv.de). In addition to these reference corpora an opportunistic corpus of ten influential national daily and weekly newspapers has been compiled. It currently consists of 4 billion tokens in 8.9 million documents. Finally several special corpora have been compiled with a total of 200 million tokens, including a large blog corpus, a corpus of contemporary interviews and a corpus of subtitles. All corpora including the historical corpora are lemmatized and annotated with the STTS tag-set (Jurish 2010). Approximately one third of the corpora is publicly available on the website of the project, the access to the other two thirds is restricted to internal use only due to copyright issues.

We distinguish the following cases that will be illustrated by some examples. Table 1 lists the number of hits that these examples generate in the DWDS corpus.

Morpho-syntactic Restrictions

Example (9) contains the nucleus [*kalter Kaffee*]$_{NP}$ (cold coffee). The absence of the morphological expansion operator ($) means that no inflected forms of the adjective and the noun are allowed. By []$_{NP}$ we denote that the two words are

Table 1: Corpus queries and corpus hits (without POS-labels for better readability)

id	query string	hits
(9)	"kalter Kaffee"	505
(9)	"$kalter Kaffee"	871
(10)	NEAR($wissen,"der Henker", 10)	48
(10.1)	"weiß der Henker"	22
(10.2)	"das weiß der Henker"	0
(12)	NEAR(ausgestorben, $sein, 10)	7273
(12)	"wie ausgestorben"	1400
(14)	NEAR("für sein Leben gern", $tun, 10)	20
(14)	"für sein Leben gern"	655
(16)	"ein Blickfang" && $sein	1613
(16)	"als Blickfang" && $dienen	153
(16)	"im Blickfang" && $stehen	0
(17)	"Bruder Leichtfuß"	287
(18)	NEAR("Blinde mit dem Krückstock", fühlen, 10)	0
(19)	NEAR("Blinde mit dem Krückstock", sehen, 10)	14

adjacent and in a noun phrase, i.e. no modifiers are allowed between the adjective and the noun. Morpho-syntactic restrictions help to reduce the number of false positives in the result sets of corpus queries. For instance the inflected forms of the adjective *kalt, kalte, kalten* would produce corpus results that do not correspond to the set phrase *kalter Kaffee*. Moreover there are optional words in (9) which we don't use for the formulation of a corpus query.

Canonical Word Order and Corpus Queries

In many cases the order of phrasal components in German set phrases is flexible. In the case of noun-verb patterns verbs can either be placed at the end of the sentence or in front of the noun. However, there are exceptions to this like example (10) where the expression has to be queried in the canonical order.

Mandatory and Optional Elements

Schemann considers the conjunction *wie* (how) to be optional in cases like (12). However, a closer look at corpora reveals that *wie* is mandatory for the phrasal meaning whereas the auxiliary verb *sein* (to be) is not, as (13) shows. In (14) the

verb *tun* (to do) is considered mandatory in the dictionary. However, there are typical example sentences without the verb, as can be seen in (15).

(12) (wie) **ausgestorben sein**
(like) extincted be
to be deserted.

(12.1) [wie|KON ausgestorben|VVPP]$_{CL}$ $sein|VAFIN

(13) In der Mittagszeit wirkt Appenzell wie ausgestorben
During lunchtime Appenzell seems deserted.

(14) (etw.) **für sein Leben gern tun**
(sth) for his life like to do
to be mad about doing sth.

(14.1) [@für|APPR) $sein|PPOSAT Leben|NN gern|ADV]$_{CL}$ $tun|VVFIN

(15) Hobbits . . . essen und trinken für ihr Leben gern und werden sehr alt. – (Die Zeit, 13.12.2012, Nr. 51)
Hobbits are mad about eating and drinking and they get very old.

Variation of the Entry and Corpus Frequency

In cases where more than one pattern exists for a given entry there can be significant differences in corpus frequency. (16) is such an entry where the variants with *sein, dienen* and *stehen* occur 1613, 153 and 0 times, respectively, in the DWDS corpus.

(16) **der/ein Blickfang sein / als Blickfang dienen / im Blickfang stehen**
To be an eyecatcher, serve as an eyecatcher, in eyecatcher stand

(16.1) [der|ARTDEF Blickfang/NN]$_{NP}$ $sein|VAFIN

(16.2) [ein|ARTINDEF Blickfang/NN]$_{NP}$ $sein|VAFIN

(16.3) [als|APPR Blickfang/NN]$_{PP}$ $dienen|VVFIN

(16.4) [im|APPRART Blickfang/NN]$_{NP}$ $stehen|VVFIN

Missing in Schemann but present in our corpora are combinations with the following verbs: *abgeben* (give away), *bieten* (to offer), *bilden* (to form), *schaffen* (to create).

Diasystematic Markers

The dictionary marks certain entries as rare, obsolete or as being related to a specific domain. This type of information can give important insights for the language learner. Corpus queries can help either to affirm the assertions made by Schemann or to correct these markers. With respect to corpora, rareness corresponds to low frequency and time-lines are useful to validate whether an expression is obsolete. Both, corpus frequency and positive time-lines, can easily be checked automatically. There are cases in which corpus evidence

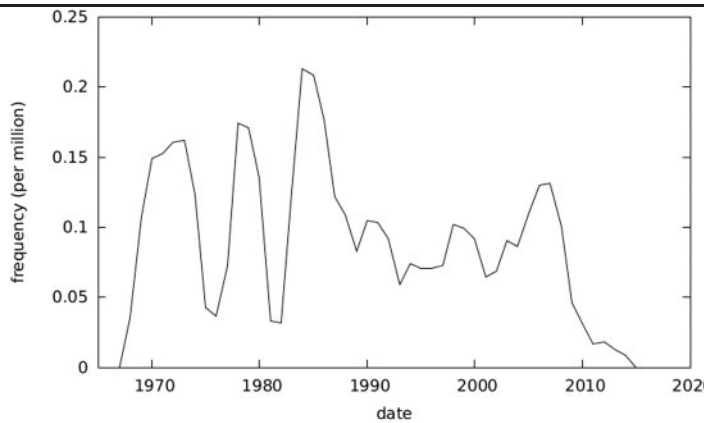

Figure 3: Evidence for the entry *Bruder Leichtfuß* in the 4 billion word corpus (DWDS)

overwrites the author's intuition. An example for this is the expression in (17), which is marked as outdated by Schemann. However, the time-line in the DWDS corpus for this expression shows sufficient evidence from 1950 until now (Figure 3).

(17) Bruder Leichtfuß
brother light foot
happy-go-lucky chap

Unnatural Contexts

The entry (18) in the dictionary does not yield any query results. The correct verb is *sehen* (to see) (19) instead of *fühlen* (to feel) because this set phrase is a play on words with *blind* and *to see*, a contradiction that would disappear with *blind* and *feel*.

(18) das fühlt (doch) (selbst) der Blinde mit dem Krückstock
this feels (still) (self) the blind with the cane

(19) das sieht (doch) (selbst) der Blinde mit dem Krückstock
this sees (still) (self) the blind with the cane
You can tell that a mile off.

(19.1) (das|PIV) $sehen|VVINF (doch) (selbst) [der|Artdef Blinde|NN mit|APPR dem|ARTDEF Krückstock|NN] $_{NP}$

6. Discussion and Future Work

In this article we presented two main results. First, the 33,237 entries in Schemann's dictionary of idioms were transformed into 43,337

lexicogrammatical patterns (section 3) which in turn correspond to 8,775 distinct pattern classes (section 4). The large number of pattern classes provides additional evidence for the difficulties corpus-based approaches have in establishing a typology based on syntactic patterns that distinguishes free grammatical patterns from phraseological patterns.

Second, our study illustrates that the collection in Schemann's dictionary greatly benefits from its validation against corpus evidence. In section 5 we provided evidence for the discrepancy between the entry structure in the dictionary and corpus evidence. There are cases where words are qualified as optional in the dictionary entry even though they are mandatory according to the corpora or, conversely, mandatory elements in the dictionary entry are optional or even not found at all in the corpora. Furthermore, in the case of alternatives marked in bold face, the dictionary lists mutually exclusive variants of an entry as being equally important even though their respective corpus frequency can vary considerably. Finally, diasystematic markers including 'obsolete' or 'rare' need to be revised in numerous cases on the basis of corpus evidence.

Very large corpora are required for the analysis of set phrases. Moon (2007) shows that even very common set phrases like *spill the beans*, *bury the hatchet* or *red herring* occur with frequencies well below one per million tokens. Moon states that the 25,000 most frequent words in English all have frequencies above one per million and she concludes with the hypothesis that set phrases are comparatively marginal items in the English lexicon. A corpus-based study with a sample of 50 set phrases on the basis of a one billion token corpus of German provides similar results (Geyken et al. 2004): the 10 most frequent set phrases of this sample occur with a frequency over 0.1 per million tokens but less than 1 per million, the ten set phrases with the lowest frequency occur less than 0.01 per million in the corpus even though each of them occurs at least once. Presumably corpora of a size well above one billion tokens are necessary for an empirically more reliable investigation of set phrases. Evidence for this assumption can be found in Hvelplund et al. (2013). In their work, a random sample of 250 low-frequency headwords of the Oxford Advanced Learner's Dictionary was chosen. Their aim was to augment these words with collocations extracted by the word sketch engine (Kilgarriff et al. 2004). They found that the 1.3 billion word corpus they used initially (UKWaC corpus, Baroni et al. 2009) turned out to be too small since they considered a collocate of less than five hits as not trustworthy, and many of the words did not have collocates above this threshold. In a second experiment a much larger 11.2 billion word corpus was used (enTenTen12). This corpus contained enough collocation data for 231 out of the 250 word sample.

The next step of our work is to compile a database of suitable corpus examples for the dictionary entries. The starting point is to convert all lexicogrammatical patterns (derived from the dictionary entries) semi-automatically

into corpus queries. More precisely, we intend to formulate corpus queries for each of the 8,775 pattern classes. There is an 'm:n' correspondence between dictionary entry and pattern class. On the one hand a pattern class contains 'n' different lexicogrammatical patterns. On the other hand a dictionary entry consists of 'm' different elementary entries (or lexicogrammatical patterns since there is a 1:1 correspondence between both) and can thus be part of more than one pattern class. We will proceed incrementally by pattern class, generating corpus queries that work as approximations for all lexicogrammatical patterns in each pattern class. Thus a corpus query for a given dictionary entry corresponds to the set of all lexicogrammatic patterns in all pattern classes that pertain to this entry. The queries will be run against the four billion word corpus of the DWDS and the query results will be evaluated for a subset (one or several lexicogrammatical patterns per pattern class) by a lexicographer in order to optimize precision and recall for each pattern class.

Even though this method sounds straightforward, there remain a few challenges. First, the generated corpus queries could yield result sets in the corpora that are too large for manual inspection. As stated above, common set phrases may have a relative frequency of under 1 per million but above 0.1 per million, these relative frequencies would correspond to result sets of 400 to 4,000 hits in a four billion word corpus. A widely used method for coping with this problem of quantity is GDEX (Good Example Extractor, Kilgarriff 2008), a software tool that suggests 'good' corpus examples to the lexicographer according to predefined criteria. In the variant of GDEX used for German (Didakowski et al. 2012) global and local criteria are distinguished. Global criteria guarantee that corpus examples are balanced with respect to text type and date of publication. Local criteria deal with correctness and comprehensibility. For example, sentences are preferred if the headword is in the main clause, shorter sentences are preferred to longer ones, sentences without free pronouns are preferred to sentences with free pronouns. The goal of GDEX is to reduce the number of corpus examples to be inspected by extracting only the n-'best' examples. Another issue worth mentioning is related to the very broad characterization of idioms in Schemann's dictionary. Substantial parts of Schemann's dictionary are collocations 'in a broad sense'. Here, alternative extraction methods based on statistics combined with syntactically parsed corpora (such as the sketch engine, Didakowski and Geyken (2012) with an implementation for German) are more effective than the mere formulation of corpus queries. Third, a practical consideration concerns the size of the resulting database. Evaluating the usefulness and lexicographic quality of all example sentences extracted from the corpora will exceed the possibilities of a small project. To this end we plan to use a collaborative method, i.e. sentences will be presented to the users who are asked to rate the appropriateness of the examples. Automatic extraction methods are becoming increasingly common in academic lexicography. It is presumed that rating automatically extracted

material is more efficient than carrying out data selection manually from scratch. This idea has been introduced under the name 'tick-box lexicography' (Rundell and Kilgarriff 2011). A practical crowd-sourcing approach for this idea was applied by Kosem et al. (2013) who describes experiments with dictionary users rating corpus examples. The platform to be used for our collaborative approach is the DWDS website, an aggregated word information system that draws on several complementary resources, including four legacy dictionaries with more than 450,000 different headwords, word statistics and corpora (www.dwds.de). With more than 35,000 registered users and 50,000 daily visits the DWDS-website is one of the most visited academic dictionary websites in Germany and as such well suited for the task at hand.

The outcome of the project outlined here will be a continuously growing database of user validated corpus examples of set phrases. This database of 'good example sentences' will not only be attractive for users of the DWDS web platform but will also constitute an ideal supplement for users of Schemann's bilingual idiom dictionaries of English, Italian, Portuguese and Spanish.

Notes

1 In the tag-set NA stands for a noun phrase placeholder such as sb. or sth.; Ptk refers to any syntactic category in the dictionary entries that cannot be subsumed under the other labels: adjective (adj), adverb (adv), conjunction (conj), determiner (det), pronoun (pron) and verb (v).

2 The author wishes to thank his former colleagues Christiane Bohn, Jordan Boyd-Graber, Rita Finkbeiner, Anastasia Chichlova, Christiane Hümmer, Kerstin Krell and Ekaterini Stathi for their contribution to transforming dictionary entries into lexico-grammatical patterns.

References

A. Dictionaries

Duden 1992. *Duden, Redewendungen und sprichwörtliche Redensarten. Wörterbuch der deutschen Idiomatik*. Mannheim: Dudenverlag, 1st edition.

Duden 2013. *Duden, Redewendungen und sprichwörtliche Redensarten. Wörterbuch der deutschen Idiomatik*. Mannheim: Dudenverlag, 4th edition.

Schemann, H. 1993. *Deutsche Idiomatik. Die deutschen Redewendungen im Kontext*. Stuttgart: Pons.

Schemann, H. 2011. *Deutsche Idiomatik. Wörterbuch der deutschen Redensarten*. Berlin / New York: de Gruyter.

B. Other literature

Baroni, M., S. Bernardini, A. Ferraresi and E. Zanchetta. 2009. 'The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora.' *Language Resources and Evaluation* 43(3): 209–226

Burger, H. 2007. 'Semantic aspects of phrasemes' In Burger, H., D. Dobrovol'skij, P. Kühn and N. R. Norrick (eds). *Phraseology. An International Handbook of Contemporary Research*. Berlin/New York: Mouton de Gruyter, 90–110.

Didakowski, J. and A. Geyken. 2012. 'From DWDS corpora to a German Word Profile – methodological problems and solutions' In Network Strategies, Access Structures and Automatic Extraction of Lexicographical Information. 2. Arbeitsbericht des wissenschaftlichen Netzwerks „Internetlexikografie". *Arbeiten zur Linguistik* 2/2012 (*OPAL*), Mannheim: Institut für deutsche Sprache, 43–52.

Didakowski, J., L. Lemnitzer and A. Geyken. 2012. 'Automatic example sentence extraction for a contemporary German dictionary' In Fjeld, R. V. and J. M. Torjusen (eds). *Proceedings of the fifteenth EURALEX International Congress*. Oslo: University of Oslo, 343–349.

Fellbaum, C. (ed.). 2007. *Collocations and Idioms: Corpus-Based Linguistic and Lexicographic Studies*. Birmingham: Continuum Press.

Firth, J. R. 1957. 'A synopsis of linguistic theory 1930-55' In *Studies in Linguistic Analysis*. Oxford: Philological Society. Reprinted in F. Palmer (ed.). 1968, Selected Papers of J. R. Firth. Harlow: Longman.

Frazer, B. 1970. 'Idioms within a transformational grammar.' *Foundations of Language* 6, 22–42.

Geyken, A. 2007. 'A reference corpus for the German language of the 20th century' In Fellbaum, C. (ed.). *Collocations and Idioms: Linguistic, Lexicographic, and Computational Aspects*. London: Continuum Press, 23–40.

Geyken A. and J. Boyd-Graber. 2003. 'Automatic classification of multi-word expressions in print dictionaries.' *Linguisticae Investigationes*, 26./2, (2003), 187–202.

Geyken, A., A. Sokirko, I. Rehbein and C. Fellbaum. 2004. 'What is the optimal corpus size for the study of idioms?' *DGfS-Jahrestagung*, Mainz 25.-27.02.2004.

Halliday, M. A. K. 1966. 'Lexis as a linguistic level' In Bazell C. E. et al. (eds). In *Memory of J.R. Firth*. London: Longman, 150–161.

Hvelplund, H., A. Kilgarriff, V. Lannoy and P. White. 2013. 'Augmenting online dictionary entries with corpus data for Search Engine Optimisation' In *Proceedings of eLex*, 2013, 66–75.

Jurish, B. 2010. 'More than words: using token context to improve canonicalization of historical German.' *Journal for Language Technology and Computational Linguistics* 25(1): 23–40.

Kilgarriff, A., M. Husák, K. McAdam, M. Rundell and P. Rychlý. 2008. 'GDEX: Automatically Finding Good Dictionary Examples in a Corpus'. In Bernal, E. and J. DeCesaris (eds). *Proceedings of the thirteenth EURALEX International Congress*. Barcelona: Universitat Pompeu Fabra, 425–433.

Kilgarriff, A., P. Rychly, P. Smrz and D. Tugwell. 2004. 'The Sketch Engine' In Williams, G. and S. Vessier (eds). *Proceedings of the Eleventh EURALEX International Congress*. Lorient: Université De Bretagne Sud, 105–116.

Kosem, I., P. Gantar and S. Krek. 2013. 'Automation of lexicographic work: an opportunity for both lexicographers and crowd-sourcing.' In Kosem, I., J. Kallas, P. Gantar, S. Krek, M. Langemets and M. Tuulik (eds). Electronic lexicography in the 21st century: thinking outside the paper. *Proceedings of the eLex 2013 conference*, 17.-19. Okt. 2013, Tallinn, Estland. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut.

Mellado Blanco, C. 2014. 'Die pragmatische Orientierung der Idiomatik Deutsch-Spanisch (2013): eine Brücke zwischen Metaphraseografie und Phraseografie' In *Vec'glav'vec've. Frazeologija in paremiologija v slovarju in vsakdanji rabi*. Jesenšek, V. and S. Babic (eds). Maribor: Maribor University, 258–274.

Moon, R. 2007. 'Corpus linguistic approaches with English corpora' In Burger, H., D. Dobrovol'skij, P. Kühn and N. R. Norrick (eds). *Phraseology. An International Handbook of Contemporary Research*. Berlin/New York: Mouton de Gruyter, 1045–1059.

Rundell, M. and A. Kilgarriff. 2011. 'Automating the creation of dictionaries: where will it all end?' In Meunier, F. et al. (eds). *A Taste for Corpora. A Tribute to Professor Sylviane Granger*. Amsterdam: Benjamins, 257–281.

Schiller, A., S. Teufel, C. Stöckert and C. Thielen. 1995. *Vorläufige Guidelines für das Taggen deutscher Textcorpora mit STTS*. Technical report, IMS, Univ. Stuttgart and SfS, Univ. Tübingen.

Sinclair, J. 1966. 'Beginning the study of lexis' In Bazell, C. E., J. C. Catford, M.A.K. Halliday and R.H.Robins. *In memory of J.R. Firth*. London: Longman, 410–430.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Tschichold, C. 2000. *Multi-word Units in Natural Language Processing*. Zürich: Olms.

Appendix

Appendix 1: Top 25 headwords ranked by entry frequency

rank	headword	translation	number of entries
1	machen	to make	291
2	Kopf	head	250
3	Hand	hand	234
4	Auge	eye	212
5	werden	to become	194
6	weg	away	187
7	leben	to live	164
8	kommen	to come	161
9	Wort	word	155
10	Herz	heart	150
11	Zeit	time	138
12	Ohr	ear	124
13	Seite	page	122
14	Welt	world	106
15	Sache	matter	99
16	Mann	man	98
17	Tod	death	98
18	Tag	day	94
19	Sinn	sense	93
20	Gesicht	face	90
21	Luft	air	88
22	Gott	god	88
23	Nase	nose	87
24	Hände	hands	85
25	sagen	to say	84

Appendix 2: Top 10 patterns ranked by entry frequency

Rank	Pattern type	frequency
1	NA NN V	659
2	NA ARTDEF NN V	621
3	NA APPR ARTDEF NN V	586
4	NA APPR NN V	513
5	NA NA APPR ARTDEF NN V	477
6	NA APPRART NN V	474
7	ARTINDEF ADJ NN	429
8	NA ARTINDEF NN V	427
9	NA ARTINDEF ADJ NN VK	409
10	NA NA APPR NN V	397
...		
8775	NA PINEG NN APPR NA APPRART NN VVINF VM	1