Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 127–134
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Collocations in Sign Language Lexicography: Towards Semantic Abstractions for Word Sense Discrimination

Gabriele Langer, Marc Schulder
Institute for German Sign Language
University of Hamburg, Germany
gabriele.langer@uni-hamburg.de, marc.schulder@uni-hamburg.de

Abstract

In general monolingual lexicography a corpus-based approach to word sense discrimination (WSD) is the current standard. Automatically generated lexical profiles such as Word Sketches provide an overview on typical uses in the form of collocate lists grouped by their part of speech categories and their syntactic dependency relations to the base item. Collocates are sorted by their typicality according to frequency-based rankings. With the advancement of sign language (SL) corpora, SL lexicography can finally be based on actual language use as reflected in corpus data. In order to use such data effectively and gain new insights on sign usage, automatically generated collocation profiles need to be developed under the special conditions and circumstances of the SL data available. One of these conditions is that many of the prerequesites for the automatic syntactic parsing of corpora are not yet available for SL. In this article we describe a collocation summary generated from *DGS Corpus* data which is used for WSD as well as in entry-writing. The summary works based on the glosses used for lemmatisation. In addition, we explore how other resources can be utilised to add an additional layer of semantic grouping to the collocation analysis. For this experimental approach we use glosses, concepts, and wordnet supersenses.

Keywords: collocations, sign language lexicography, corpus-based analysis of sign usage, lexical profile, word sense discrimination, sign language NLP, cross-lingual bootstrapping

1. Introduction

One central task in the lexicographic analysis of a language is to identify different word senses, i. e. different uses of a given word or expression.[1] In corpus-based lexicography, this is achieved by inspecting many occurrences of the word of interest in their linguistic context to determine their contextual meanings and conditions of use. Occurrences of same or similar meanings and usages are then grouped together as representing a specific use of the expression type. Each use that is identified in this manner is then described as a particular sense in the dictionary entry. It is necessary to consider as many examples of usage as possible in order to identify and substantiate the existence of specific word senses and prevent overlooking typical uses.

During lexicographic analysis, single occurrences and their immediate linguistic context are usually reviewed in a concordance view of the data (also known as *Keyword in Context* or *KWIC*). However, for high frequency words in large corpora it is impossible to inspect all existing occurrences manually. Instead, lexicographers either resort to inspecting a randomised sample of occurrences from different sources (Landau, 2001, p. 296) or to consulting the output of a lexical profiling software. Lexical profiles such as the Word Sketches generated by Sketch Engine are summaries of collocate lists grouped by part of speech (POS) tag sequences. As “[...] the regular lexical environment of a word is one of the most reliable indicators if its senses”, they provide lexicographers with a “[...] preferred starting point to their analyses of complex headwords.” (Atkins and Rundell, 2008, p. 110-111).

In sign language linguistics, corpus-based lexicography is still a very new field. In general, sign languages (SL) belong to the less researched and less resourced languages and are without a written form and tradition. This presents technical and methodological challenges when collecting, annotating and analysing SL texts in a corpus and when describing signs and their uses in dictionaries (Zwitserlood et al., 2013). For instance they lack sufficient numbers of sources to sample from, as well as most natural language processing (NLP) tools, such as automatic POS taggers.

To enable corpus-based lexicographic work and research on usage in SL it would be very useful to have concordance views and lexical profiles with collocational information of the target sign. Without a direct written representation of signing, ways of representing, grouping and ranking occurrences must be found that do not rely on an unmediated written representation of the language samples under investigation. In SL corpus linguistics the most feasible way is to use the gloss annotation for all aspects of searching, sorting and counting while time-aligned video files stay easily accessible for viewing the actual language samples in their direct, unmediated form (cf. Johnston, 2010).

In this article we explore such an approach. Section 2 provides relevant background on sign language linguistics, while Section 3 introduces the corpus data used in our work. Section 4 gives general information on collocations. In Section 5 we introduce a collocation view based on sign glosses, while in Section 6 we explore the use of automatic cross-lingual methods for providing an additional semantic collocation layer, based on wordnet supersenses. Section 7 provides an outlook on possible future extensions of our work.

[1] For the purpose of brevity, when talking about abstractions that refer to both signed and spoken languages we use terminology from lexicography and linguistics which in some cases reflects the spoken language bias of its origin. When referring to the general concept of a lexical unit, the term *word* covers both the word of a spoken language and the sign of a signed language. Similarly, *word sense* also covers *sign sense* and *phonological variant* covers *cherological variant*.

2. Background

In this section we provide a brief overview of sign language research, specifically regarding language transcription (Section 2.1), the state of research on general linguistic concepts (Section 2.2) and the availability of natural language processing software (Section 2.3).

2.1. Transcription of Sign Languages

Since sign languages lack a standard form of written representation and the usual writing systems are not suitable to represent signing, it is a common method in SL research to represent signing in a stable, non-perishing form via the use of glosses. In SL corpus research a gloss is a metalinguistic label to represent a specific sign type. Each sign type is assigned a unique gloss that is used consistently and exclusively for that sign (Johnston, 2010). A gloss is comprised of a spoken language word (the *gloss name*, usually written in capital letters) and additional suffixes such as digits and letters to distinguish between different signs and variants.

The purpose of the gloss as a label is that it can be written, read and easily remembered by humans as well as sorted, counted and manipulated by computers.[2] Through the representation of sign texts as glosses certain NLP operations can be performed on these representations. However, glosses by themselves do not reveal anything about the actual sign forms. For this they need to be related and (preferably) directly linked to a corresponding lexical entry. Furthermore, glosses are not to be mistaken as context-appropriate translations of a sign. Using a spoken language word as a label for a sign bears the risk of unsuitably inferring that syntactic and semantic properties of the spoken language word, such as parts of speech and nuances of meaning, also are valid for the sign of the sign language (cf. Slobin, 2008).

Bearing these drawbacks and risks in mind, gloss names can nevertheless be used for NLP purposes as a general, rough and incomplete approximation to a sign's general meaning, as gloss names are usually chosen so that the spoken language word indicates a core meaning of the respective signs (see, for example, Johnston (2010, pp. 119–120)).

2.2. The State of Sign Language Research

One of the major challenges of linguistic research for sign languages is that there still are no commonly agreed upon theories for certain basic categories and concepts, such as sentences and parts of speech. As languages without a written form and primarily used in face-to-face communication, sign languages share some structural properties with spoken language mode (as opposed to written language), for example the difficult issue of determining sentence boundaries. While written language is usually segmented based on orthographic markers, the spoken language mode of spoken languages has been very difficult to segment (Auer, 2010; Westpfahl and Gorisch, 2018) and thus segmentation in many speech corpora is based solely on pauses in speech (Hamaker et al., 1998; Schmidt, 2014). In addition, the visual modality of SL informs very different language structures that still lack comprehensive analysis and description. As Schwager and Zeshan (2008) observe, "[b]y and large, it has not been easy to identify workable syntactic tests for sign languages, given that they often have relatively free word order and some of their sentence structures are unfamiliar from a spoken language background, including spatial syntax and simultaneous constructions." For instance, in German Sign Language many signs can be multi-functional, appearing for example as either predicate or argument. The body of research does not yet provide a commonly agreed background on sentence structures and POS categories that can robustly be used for segmenting sentences and tag signs for POS in corpus annotation.

2.3. NLP for Sign Languages

Presently, sign language research has to manage without most of the NLP tools and procedures that are ubiquitous for well researched and resourced spoken languages like English or German. Apart from the general struggles of any less resourced language, such as a lack of (machine-readable) language data, sign language research is made especially challenging by its specific modality in the visual-gestural domain, the resulting very different language structures, the lack of comprehensive grammars that categorise and describe these structures and could be used for widespread analyses, and the absence of large machine-readable lexical resources.

One reason for the lack of NLP tools for sign languages is that for almost any NLP task, a series of other tasks must be executed in preparation for it. The most common of these pre-processing steps are a) sentence tokenisation (splitting a text into sentences), b) word tokenisation (splitting a sentence into words or signs), c) part of speech tagging, and d) lemmatisation (turning a word or sign into its citation form). Even these steps build upon one another (e. g. the lemma of a word depends on its part of speech). As we discussed in Section 2.2, neither sentence boundaries nor parts of speech are even fully defined for sign languages, so designing machine classifiers for them (let alone for more advanced tasks) is not yet feasible. NLP for sign language research therefore mostly consists to makeshift solutions, broad generalisations and solutions bootstrapped from spoken language data and tools.

3. Data for German Sign Language

3.1. The DGS Corpus

For lexicographic analyses and descriptions, such as the collocation analyses discussed in this article, we exclusively use the data of the *DGS Corpus*. This corpus of filmed natural and near-natural conversations and narrations signed in German Sign Language (DGS) has been collected in Germany between 2010 and 2012 (Nishio et al., 2010). It includes language samples of 330 persons from all over Germany who were filmed in pairs. 560 hours of signing were recorded, of which about 79 hrs of running text have been annotated and lemmatised in iLex so far. As of February 2020, the corpus contains close to 530,000 token tags. About 50 hours of the material have been published as the *Public DGS Corpus* and are available online

[2] In the case of iLex, instead of using unique glosses the iLex-internal unique ids can also be used for identification, counting and other computational tasks. However, in the following we describe the approach as based on the glosses.

through the community portal *MY DGS*[3] and the research portal *MY DGS – annotated*[4]. Most lemmatised running texts in the *DGS Corpus* data are translated into German and time-aligned with utterances as suggested by the German translation. For collocation analyses and lexicographic description all lemmatised corpus data is used, including lemmatised data of the unpublished material.

3.2. The iLex Database of the IDGS

The work described in this article is conducted in the environment of iLex, an integrated annotation tool and lexical database (Hanke and Storz, 2008). During lemmatisation tokens are matched to types that are defined as lexical entries. A token can be lemmatised only after the respective type has first been established as a lexical entry. Thus, annotation in iLex consequently leads to the identification of sign types and their description in lexical entries.

At the Institute of German Sign Language and Communication of the Deaf (IDGS) at Hamburg University we use a complex structure of **types** and subordinate **subtypes** for token-type matching during lemmatisation called *double glossing* (Langer et al., 2016; Konrad et al., 2018).[5]

A type is an abstract unit of the sign language with a specific form often associated with an underlying image – that can have several differing realisations in actual use – and with one or several conventional meanings. Each subtype belongs to a specific type and roughly represents one of its conventional uses. The conventional meaning that a subtype covers is described by linking one or more **concepts** to the particular subtype entry. *Concepts* here are to be understood as pre-theoretical and pre-analytical indications of conventional meanings. In iLex they are their own entities that are identified by a German word or expression covering the conventional meaning of the sign. When the German word is ambiguous, a disambiguating **context** can be added to specify the meaning.

Often the words that are chosen as concept descriptions to indicate a conventional meaning of the subtype correspond to a specific mouthing associated with the sign. For signs that are presumed to be multifunctional, several concepts for the same German root may be specified to represent their POS-specific forms (e. g. *"Arbeit"* (noun) and *"arbeiten"* (verb)).

During lemmatisation sign tokens are matched to a type or one of its subtypes. Each type and subtype is identified by a internal unique id, a citation form described in HamNoSys notation (Hanke, 2004), and – for the benefit of the human user – receives a unique gloss. iLex ensures that glosses for types and subtypes are unique.

The iLex database used and maintained by the IDGS comprises data from several projects. Data from all included projects commonly use the same type and subtype entries for lemmatisation. The DGS-Korpus project (cf. Section 3.1) re-uses the sign type entries established in previous projects with their descriptions as a starting point and adds, re-evaluates and corrects them when necessary.

We exclusively use tokens from the *DGS Corpus* data for collocational analyses, but the sign type entry information – namely gloss names and concepts – may stem from previous projects. As the linking of concepts to subtypes was done primarily on the basis of introspection and in accordance with the different guidelines of these projects, this information can be of varying quality.

The basic annotation of the *DGS Corpus* does not attempt to identify and tag segments of DGS other than individual signs. The only available approximation to DGS sentences or utterances are the German translation tags. German translations are source-language oriented translations, but they still constitute sentences or utterances in German. As part of the annotation they are time-aligned to the DGS signing in the video and token tags in the transcript. Segment lengths were determined pragmatically while as many of the DGS structures and boundary signals as possible were taken into consideration (cf. Section *Translation into German* in Konrad et al. (2018)). Needless to say that this is not an ideal substitute for monolingual structure-based segmentation and only a rough approximation, as sentence structures in the German translations and structures in the DGS source text may differ to a some extent from each other.

For the purpose of this article we refer to two data types that are available to us: glosses and concepts.

4. Collocation

One of the advantages of a corpus is that collocational information can be extracted by statistical means from corpus data. Having a relatively large SL corpus available allows for new insights on sign usage and meaning and thus also for new kinds of information on signs in dictionary entries. In the context of our lexicographic work, collocational information is relevant in the following ways:

1. Collocational information is a good indicator of different word senses and can be utilised to support lexicographers in word sense discrimination (WSD).

2. Information on typical sign combinations and other usage patterns is information to be included in dictionary entries because it is useful especially for language learners. Lists of frequent neighbours of the target signs suggest candidates for collocations, phrases and (loan) compounds to be included in the dictionary entry of the *DW-DGS*.[6]

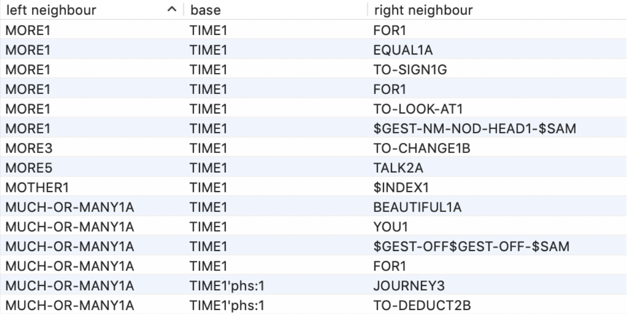
In this article we focus on methods of extracting collocational information from the *DGS Corpus* to support WSD. Considering that a) the investigation of collocations in sign language is only just becoming possible thanks to recent advances in SL corpus creation, b) DGS exhibits considerable phonological and lexical variation, and c) borders between individual signs and their variants might be less clear than their categorisation via glosses suggests, we adopt a rather pragmatic and broad definition of collocation for the time

[3] `http://meine-dgs.de`

[4] `http://ling.meine-dgs.de`

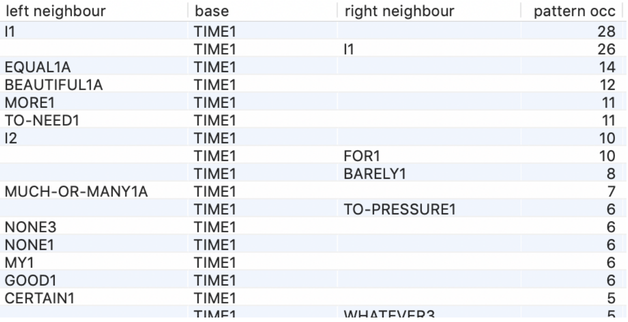
[5] For the purposes of this article we simplify the type structure slightly. For a discussion of the complete type structure, see Langer et al. (2018).

[6] `http://dw-dgs.meine-dgs.de`



left neighbour ^	base	right neighbour
MORE1	TIME1	FOR1
MORE1	TIME1	EQUAL1A
MORE1	TIME1	TO-SIGN1G
MORE1	TIME1	FOR1
MORE1	TIME1	TO-LOOK-AT1
MORE1	TIME1	$GEST-NM-NOD-HEAD1-$SAM
MORE3	TIME1	TO-CHANGE1B
MORE5	TIME1	TALK2A
MOTHER1	TIME1	$INDEX1
MUCH-OR-MANY1A	TIME1	BEAUTIFUL1A
MUCH-OR-MANY1A	TIME1	YOU1
MUCH-OR-MANY1A	TIME1	$GEST-OFF$GEST-OFF-$SAM
MUCH-OR-MANY1A	TIME1	FOR1
MUCH-OR-MANY1A	TIME1'phs:1	JOURNEY3
MUCH-OR-MANY1A	TIME1'phs:1	TO-DEDUCT2B

Figure 1: Excerpt of a list of 525 tokens of type TIME1 with left and right neighbour glosses. Occurrences are sorted by the alphabetical order of the left neighbour.



left neighbour	base	right neighbour	pattern occ
I1	TIME1		28
	TIME1	I1	26
EQUAL1A	TIME1		14
BEAUTIFUL1A	TIME1		12
MORE1	TIME1		11
TO-NEED1	TIME1		11
I2	TIME1		10
	TIME1	FOR1	10
	TIME1	BARELY1	8
MUCH-OR-MANY1A	TIME1		7
	TIME1	TO-PRESSURE1	6
NONE3	TIME1		6
NONE1	TIME1		6
MY1	TIME1		6
GOOD1	TIME1		6
CERTAIN1	TIME1		5
	TIME1	WHATEVER3	5

Figure 2: Distinct neighbours, sorted by bi-gram frequency and a minimum frequency of five or more.

being.[7] In the context of lexicographic work with the dictionary user in mind the definition presented by Fuertes-Olivera et al. (2012, p. 299) can serve as a starting point: "The term collocation was chosen as an umbrella term for referring to word combinations that are typical for the kind of language in question, and which can be useful for re-use in text production or for assisting in text translation. They are composed of two or more orthographic words, do not constitute a full sentence, but offer potential users the possibility of obtaining relevant information [...]". In our case, we consider individual (simplex) signs represented by glosses, as our data contains no orthographic words.

Collocational information in our approach includes all kinds of multi-sign patterns, especially including *collocations* in the narrow sense and *selectional restrictions*. According to Atkins and Rundell (2008, p. 302), "[. . .] [b]oth terms refer to an observable tendency of certain words to occur frequently with certain other words. When we talk about 'selectional restrictions', we mean the general semantic category of items that typically appear as the subjects or objects of a verb, or or as the complements of an adjective. A collocation on the other hand, is a recurrent combination of words, where *one specific lexical item* (the 'node') has an observable tendency to occur with another (the 'collocate'), with a frequency greater than chance."[8]

Another definition of collocation that has "received general approval among lexicographers" (Orlandi, 2016, p. 26) is that of Bartsch (2004, p. 76) where collocations are "lexically and/or pragmatically constrained recurrent co-occurrences of at least two lexical items which are in a direct syntactical relation with each other". This definition includes what fully-fledged lexical profiles make explicit, that is, taking syntactic relations (e. g. dependencies) into account when determining and presenting collocations.

A relevant question for SL lexicography is how SL corpora can be used to automatically generate comparably informative collocation profiles for signs, even when many of the prerequesites for such an automated analysis do not yet exist for SL data. It would be very useful to find methods to identify frequent neighbours and to group them with regard to their syntactic or semantic relation to the base, be it argument structure or other kinds of functional or semantic categories. Collocation analysis for *DGS Corpus* data specifically is best done at the subtype level because subtypes correspond with conventional sign uses and pre-group occurrences according to these roughly defined meanings. In the following sections we discuss our approach to NLP supported detection, grouping, and presentation of collocations for lexicographic purposes.

5. Gloss-based Collocations

A very simple first approach for identifying frequent neighbours of a target sign, shown in Figure 1, is to run a query that returns all occurrences of the target sign and its left (or right) neighbours. Results are ordered alphabetically by the neighbour gloss name. The human eye will very quickly find groups of identical neighbours in this list.

The query can be refined to provide a better overview by showing each distinct left and right neighbour only once and count and display the number of occurrences for this combination (bi-gram) in the result. In Figure 2 we see a frequent neighbours list grouped by distinct neighbour glosses and ordered by the frequency count of the bi-gram. Groups with fewer than five members are filtered out.

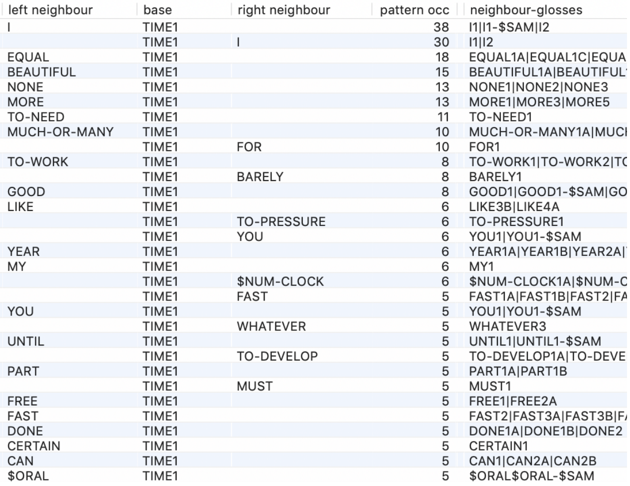
Up to this point the query shows collocations in the more narrow sense, i. e. frequent combinations of specific lexical items. However, as the corpus is still limited in size and most types have rather small numbers of tokens, relevant semantic patterns may not show up, especially since phonological variants are covered by separate types in the *DGS Corpus*.[9] For pattern detection the distinction of phonological variants is too fine-grained and variant types should be grouped together in the analysis. Furthermore, DGS not only exhibits a high amount of phonological variation, but also of lexical variation, often even within a single region or by a single individual.[10] For the purpose of WSD the

[7] For a good overview on defining collocation – especially in the field of lexicography – see Orlandi (2016).

[8] *Selectional restrictions* have also been called *selectional preference*, cf. Sinclair (1996).

[9] Phonological variants are marked by the same number but different letters after the gloss, e. g. MUCH-OR-MANY1A and MUCH-OR-MANY1B.

[10] In the *DGS Corpus* lexical variants share the same gloss name but receive different numbers, e. g. NONE1 and NONE3. In some cases distinct meanings of a German word (polysemes and homonyms) may lead to the same gloss name being used for signs with distinct form and meaning. While this is a potential source of errors, it is in practice an acceptable trade-off, as such signs are expected to occur in distinctly different collocational contexts.



left neighbour	base	right neighbour	pattern occ	neighbour-glosses
I	TIME1		38	I1|I1-$SAM|I2
	TIME1	I	30	I1|I2
EQUAL	TIME1		18	EQUAL1A|EQUAL1C|EQUA
BEAUTIFUL	TIME1		15	BEAUTIFUL1A|BEAUTIFUL
NONE	TIME1		13	NONE1|NONE2|NONE3
MORE	TIME1		13	MORE1|MORE3|MORE5
TO-NEED	TIME1		11	TO-NEED1
MUCH-OR-MANY	TIME1		10	MUCH-OR-MANY1A|MUCI
	TIME1	FOR	10	FOR1
TO-WORK	TIME1		8	TO-WORK1|TO-WORK2|TC
	TIME1	BARELY	8	BARELY1
GOOD	TIME1		8	GOOD1|GOOD1-$SAM|GO
LIKE	TIME1		6	LIKE3B|LIKE4A
	TIME1	TO-PRESSURE	6	TO-PRESSURE1
	TIME1	YOU	6	YOU1|YOU1-$SAM
YEAR	TIME1		6	YEAR1A|YEAR1B|YEAR2A|
MY	TIME1		6	MY1
	TIME1	$NUM-CLOCK	6	$NUM-CLOCK1A\|$NUM-C
	TIME1	FAST	5	FAST1A|FAST1B|FAST2|FA
YOU	TIME1		5	YOU1|YOU1-$SAM
	TIME1	WHATEVER	5	WHATEVER3
UNTIL	TIME1		5	UNTIL1|UNTIL1-$SAM
	TIME1	TO-DEVELOP	5	TO-DEVELOP1A|TO-DEVE
PART	TIME1		5	PART1A|PART1B
	TIME1	MUST	5	MUST1
FREE	TIME1		5	FREE1|FREE2A
FAST	TIME1		5	FAST2|FAST3A|FAST3B|F
DONE	TIME1		5	DONE1A|DONE1B|DONE2
CERTAIN	TIME1		5	CERTAIN1
CAN	TIME1		5	CAN1|CAN2A|CAN2B
$ORAL	TIME1		5	$ORAL$ORAL-$SAM

Figure 3: Frequent neighbours analysis similar to Figure 2, but with collapsed phonological and lexical variants. Variants included in a group are listed in its rightmost column.

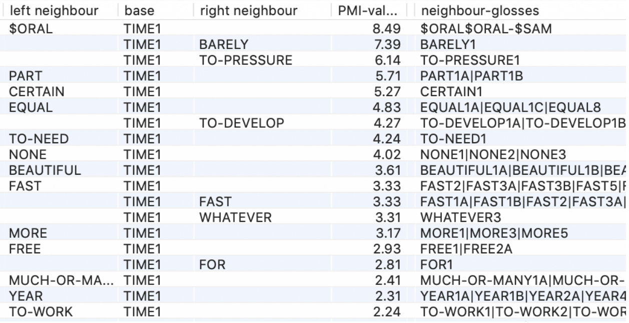
focus of interest is less on specific lexical items (forms) but on the typical semantic context the target sign is used in. Grouping lexical as well as phonological variants together in neighbourhood pattern analysis can help to identify different senses of the specific target sign (base). Thus, for neighbourhood analysis not the individual types but all types with the same gloss name are collapsed into one group to leverage the information on phonological and lexical variation as coded in the full glosses. The result of this query can be seen in Figure 3.

An advantage of this oversimplification is that more bi-gram combinations are feeding into the general semantic patterns so that more relevant patterns show up for the target sign. At this point the analysis is not covering individual collocations in the narrow sense anymore, but this level of granularity has proven fruitful for the corpus size and variant-richness of the *DGS Corpus*.

Regarding the sorting of results, there is further room for improvement. Sorting target-neighbour pairs by their raw co-occurrence frequency has an inherent bias towards signs that are generally more frequent, as they have a higher likelihood to co-occur with the target by chance without being particularly relevant. For example, in Figure 3 the gloss name `I` is ranked most highly because it covers two of the most frequent signs in the corpus, rather than due to any particular relevance for `TIME1`.

To address this bias, Church and Hanks (1990) introduced the use of **pointwise mutual information (PMI)** (Shannon, 1948) to lexicography.[11] Given the individual frequencies $f(x)$ and $f(y)$ of target and neighbour, their co-occurrence frequency $f(x, y)$ and the overall number of corpus tokens N, PMI is defined as:

$$\mathrm{I}(x, y) = \log_2 \frac{f(x, y)\, N}{f(x)\, f(y)} \tag{1}$$



left neighbour	base	right neighbour	PMI-val...	neighbour-glosses
$ORAL	TIME1		8.49	$ORAL$ORAL-$SAM
	TIME1	BARELY	7.39	BARELY1
	TIME1	TO-PRESSURE	6.14	TO-PRESSURE1
PART	TIME1		5.71	PART1A|PART1B
CERTAIN	TIME1		5.27	CERTAIN1
EQUAL	TIME1		4.83	EQUAL1A|EQUAL1C|EQUAL8
	TIME1	TO-DEVELOP	4.27	TO-DEVELOP1A|TO-DEVELOP1B
TO-NEED	TIME1		4.24	TO-NEED1
NONE	TIME1		4.02	NONE1|NONE2|NONE3
BEAUTIFUL	TIME1		3.61	BEAUTIFUL1A|BEAUTIFUL1B|BE
FAST	TIME1		3.33	FAST2|FAST3A|FAST3B|FAST5|F
	TIME1	FAST	3.33	FAST1A|FAST1B|FAST2|FAST3A|
	TIME1	WHATEVER	3.31	WHATEVER3
MORE	TIME1		3.17	MORE1|MORE3|MORE5
FREE	TIME1		2.93	FREE1|FREE2A
	TIME1	FOR	2.81	FOR1
MUCH-OR-MA...	TIME1		2.41	MUCH-OR-MANY1A|MUCH-OR-
YEAR	TIME1		2.31	YEAR1A|YEAR1B|YEAR2A|YEAR4
TO-WORK	TIME1		2.24	TO-WORK1|TO-WORK2|TO-WOR

Figure 4: Collocation list view of frequent left and right neighbours of subtypes of the type `TIME1`, ordered by PMI value of the bi-gram combination; bi-grams with fewer than five occurrences are omitted.

Figure 4 shows the neighbourhood patterns query ordered by PMI measure. This collocation list view has proved itself useful for lexicographic analysis in the DGS-Korpus project for some years now and serves as a substitute for a not yet available full-grown collocational profile of the target sign under investigation.[12]

6. Supersense Collocations

The Word Sketch profiles enhance their collocation lists by providing a semantic clustering of collocates, e. g. by grouping near-synonyms together based on information from a thesaurus (cf. Atkins and Rundell, 2008, p. 111). In this section, we explore how semantic groupings can be realised for a sign language.

6.1. The Need for Semantic Categories

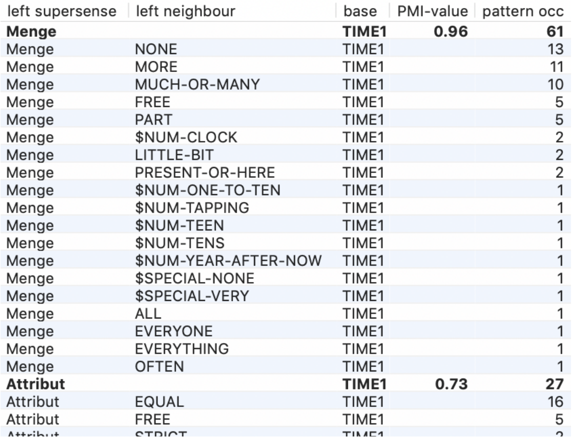
While using the collocation list view presented in Section 5 to support the analyses of sign usage, lexicographers in the DGS-Korpus project noticed wider semantic and syntactic patterns across listed neighbours. Sometimes several neighbours were identified as members of a category that could be described by an abstract criterion of semantic grouping or by a functional or presumed syntactic role. For example, several left neighbour collocates of `TIME1` are signs that have a gloss name indicating a quantifying relation with `TIME1`, e. g. `MUCH-OR-MANY`, `MORE` and `NONE`. Many left neighbour collocates of `TO-SAY1` are signs referring to persons filling the semantic role of agent as argument of `TO-SAY1`.

These patterns are often examples of semantic restriction/preference and at the same time can indicate dependency structures and syntactic functions, as these phenomena are often related (cf. Bartsch, 2004, pp. 70–71). They are very useful for WSD and also constitute valuable information on sign usage for language learners using the dictionary. Naturally, such gloss patterns must be verified against DGS data by inspecting the actual signed utterances.

However, some of these patterns may go unnoticed because each individual bi-gram contributing to the pattern may by itself be too infrequent to show up in the collocation list (see our use of frequency thresholds in Section 5). Only

[11] Lexicographers have since introduced a variety of other metrics, e. g. the *logDice* formula (Rychlý, 2008) used by *Sketch Engine*. Most of these require additional syntactic information and are therefore not suitable for our purposes (see Section 2.2).

[12] Langer et al. (2018) mention the approach, albeit in less detail, as part of their discussion of views for lexicographic work.



left supersense	left neighbour	base	PMI-value	pattern occ
Menge		**TIME1**	**0.96**	**61**
Menge	NONE	TIME1		13
Menge	MORE	TIME1		11
Menge	MUCH-OR-MANY	TIME1		10
Menge	FREE	TIME1		5
Menge	PART	TIME1		5
Menge	$NUM-CLOCK	TIME1		2
Menge	LITTLE-BIT	TIME1		2
Menge	PRESENT-OR-HERE	TIME1		2
Menge	$NUM-ONE-TO-TEN	TIME1		1
Menge	$NUM-TAPPING	TIME1		1
Menge	$NUM-TEEN	TIME1		1
Menge	$NUM-TENS	TIME1		1
Menge	$NUM-YEAR-AFTER-NOW	TIME1		1
Menge	$SPECIAL-NONE	TIME1		1
Menge	$SPECIAL-VERY	TIME1		1
Menge	ALL	TIME1		1
Menge	EVERYONE	TIME1		1
Menge	EVERYTHING	TIME1		1
Menge	OFTEN	TIME1		1
Attribut		**TIME1**	**0.73**	**27**
Attribut	EQUAL	TIME1		16
Attribut	FREE	TIME1		5
Attribut	STRICT	TIME1		2

Figure 5: Supersense collocation of TIME1. Supersenses are ranked by their PMI value. Below each supersense we show the gloss names that are part of its collocation.

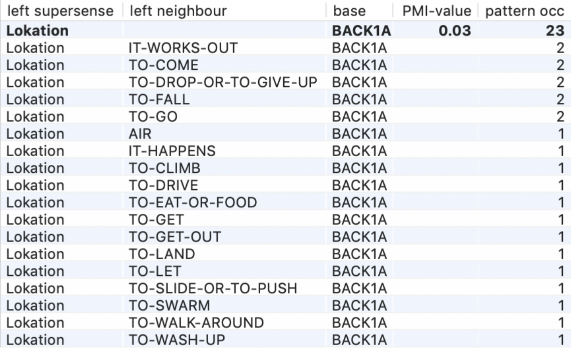
after grouping all these infrequent signs together would it become apparent how frequent the pattern that they are part of may in fact be.

This presents us with a chicken and egg problem. To find the pattern we need to see the infrequent signs and to see the infrequent signs we need to have already grouped them according to our pattern. Syntactic information like parts of speech and dependency structures that might help structure our data further are not available to us. Instead, we take inspiration from Atkins and Rundell (2008), who mention *selectional restrictions* as manifestations of usage and therefore helpful to discriminate between different senses: "When we talk about 'selectional restrictions', we mean the general semantic category of items that typically appear as the subjects or objects of a verb, or as the complements of an adjective. [...] [O]nce you know the category, any word belonging to that category can fill the relevant slot" (Atkins and Rundell, 2008, pp. 302–303). While we cannot use selectional restrictions (this would require syntactic information about parts of speech and their arguments), we might still be able to group signs into semantic categories if we can access an appropriate semantic resource.

6.2. Wordnet Supersenses

A **wordnet** is a lexical resource of semantic relations between words of a specific language (Miller et al., 1990). Words are organised by their word senses and grouped with other words of the same sense into *synsets* (synonym sets). Synsets are linked by various relations, such as hyperonymy, meronymy or entailment. Each synset is also assigned a so-called *supersense* (also known as *lexicographer sense*). Supersenses are coarse semantic categories, such as person, location or emotion. They might therefore be used as semantic categories in our list of collocations.

No wordnet exists yet for DGS, so we instead leverage German-language components of the *DGS Corpus* to extract supersenses from the German wordnet *GermaNet* (Hamp and Feldweg, 1997). As we are looking to retrieve semantic generalisations, rather than nuances, we believe this to be an acceptable compromise.



left supersense	left neighbour	base	PMI-value	pattern occ
Lokation		**BACK1A**	**0.03**	**23**
Lokation	IT-WORKS-OUT	BACK1A		2
Lokation	TO-COME	BACK1A		2
Lokation	TO-DROP-OR-TO-GIVE-UP	BACK1A		2
Lokation	TO-FALL	BACK1A		2
Lokation	TO-GO	BACK1A		2
Lokation	AIR	BACK1A		1
Lokation	IT-HAPPENS	BACK1A		1
Lokation	TO-CLIMB	BACK1A		1
Lokation	TO-DRIVE	BACK1A		1
Lokation	TO-EAT-OR-FOOD	BACK1A		1
Lokation	TO-GET	BACK1A		1
Lokation	TO-GET-OUT	BACK1A		1
Lokation	TO-LAND	BACK1A		1
Lokation	TO-LET	BACK1A		1
Lokation	TO-SLIDE-OR-TO-PUSH	BACK1A		1
Lokation	TO-SWARM	BACK1A		1
Lokation	TO-WALK-AROUND	BACK1A		1
Lokation	TO-WASH-UP	BACK1A		1

Figure 6: Excerpt of the supersense collocation of BACK1, showing the supersense collocate Lokation (location) and the gloss names it contains. None of the individual names occurs more than twice in the corpus, but grouped into the supersense the semantic pattern becomes apparent.

To connect DGS signs to *GermaNet* supersenses, we use the concept entries associated with subtypes in the *DGS Corpus* (see Section 3.2). As concept entries are (approximate) indications of the conventional meanings of a sign and are written as single German expressions, we treat them as rough German equivalents. Each sign can have several concepts, giving us a one-to-many mapping to German words. For each sign concept we look up matching terms in *GermaNet* across all parts of speech and retrieve the synsets which they are part of. The supersenses of these synsets are then treated as the supersenses of the concept. The supersenses of a sign are the set of supersenses of all concepts of that sign. Similarly, the supersenses of a gloss name group (i. e. a set of signs grouped by the name-component of their gloss, see Section 5) consist of all supersenses of its signs.

Having now bootstrapped supersense categories for our DGS sign inventory, we return to the task of creating a collocation list view. First we follow the steps of the gloss-based collocations pipeline from Section 5. The neighbouring tokens of the base are grouped by their signs to establish collocates. These sign collocates are collapsed further into gloss name collocates. Instead of then listing the gloss name collocates directly, we look up their supersenses and use them to create supersense collocates. Each supersense collocate contains every associated gloss name collocate. This means a gloss name collocate can occur in multiple supersense collocates if it has multiple supersenses. The supersense collocates are then ranked by their PMI and the list is pruned at the usual frequency threshold. In the collocation list view, each supersense collocate is followed by a list of the gloss name collocates that it is comprised of. An example of this can be seen in Figure 5.

Note that the frequency threshold only applies to the supersense collocate, not to individual gloss name collocates it contains. This allows the long tail of low-frequency collocations to still impact the ranking of the semantic categories that they are a part of. For example, Figure 6 shows an excerpt of the supersense collocation of BACK1. One sense of this sign can be described as *"moving back to a place where one came from or has been before"*. Often this sign follows other signs of movement across space, such as TO-GO-THERE, TO-COME, TO-DRIVE, TO-GET-OUT and others. However, none of these neighbour signs co-occur with BACK1 more than once or twice in the corpus. Due to these low individual frequencies, they are omitted in the gloss name collocation view. As we can see in Figure 6, the same is not true for the supersense collocation view. Here the neighbour signs are grouped together under the supersense Lokation (location), clearly showing the collocation pattern of the sign sense. Cases like this show how the supersense collocation view can be a useful addition to the lexicographer's toolkit, especially when used in concert with the gloss name collocation view.

6.3. Fallback: Glosses as Concepts

Our approach for connecting signs with *GermaNet* supersenses relies on the availability of concept entries as a bridge between languages. However, for many other sign language datasets, such an explicit cross-lingual semantic layer is not available. A possible fallback solution can be the use of gloss names as impromptu concepts. While there are obvious drawbacks to this (no multiple concepts per sign, as well as the established dangers of treating glosses as translations) it may still be an acceptable compromise to provide lexicographers with another tool in their toolbox.

On the technical side, certain complications arise as well. As glosses are commonly written in all caps, capitalisation of the host language is lost, which may create ambiguities, depending on the language (e. g. in German *laut* means 'loud', but *Laut* means 'sound'). Depending on the exact annotation guidelines used for naming glosses, a variety of multi-word gloss ambiguities may also have to be resolved. In the *DGS Corpus*, for example, hyphenation can fulfil a number of different functions. It can indicate actual multi-word expressions (ACH-SO1, *ach so*, 'I see') or fine-grained meanings that require more than a single German term to describe (ANMACHEN-BILDSCHIRM1, *anmachen (Bildschirm)*, 'turn on (monitor)'). It can be used to provide disambiguating contexts (FREI-KOSTENLOS1 means 'free' (*frei*) in the sense of 'at no charge' (*kostenlos*), but not 'unclaimed' or 'not imprisoned') or to append foreign language markers (HOLLYWOOD-ASL1 is a sign in American Sign Language). And, of course, sometimes a hyphen is simply part of a word (S-BAHN1, *S-Bahn*, 'commuter train').

6.4. Caveats and Finetuning

Using *GermaNet* supersenses to group signs is a first step towards bootstrapping semantic categories for DGS. It is not, however, without its caveats. The first one is that supersenses are extremely broad categories. Not counting duplicates across different parts of speech, *GermaNet* provides a total of 38 different supersenses. Supersense collocations can therefore only ever be a first filter, followed by a more thorough analysis.

Another problem stems from the fact that we access *GermaNet*, a primarily sense-based resource, via lemma-based lookups. As we are unable to determine ahead of time which word senses apply to the tokens in a given collocate, we are bound to overgenerate, selecting more senses than necessary and thus extracting incorrect supersenses. This issue is exacerbated by the fact that we are performing this lookup cross-lingually, thus capturing word senses of a German translation that do not apply to the sign at all. We expect that the impact of these issues is lessened in our specific case, as many of the incorrect senses will share a supersense with correct senses. Also, as the resulting output is intended for lexicographic work, any suggested patterns will be further scrutinised by the lexicographer.

The third issue is one of lexical coverage. While *GermaNet* covers a very large vocabulary, it focuses on content words, especially nouns, verbs and adjectives. Function words are omitted. It also does not cover names and has only a limited selection of location names.

To address the last two issues at least in part, we introduce a number of additional steps when determining supersenses for signs. As was mentioned in Section 3.2, concept entries can be given a disambiguating *context* when the German concept term by itself is too ambiguous (e. g. *"Bayern"* can refer to either the German federal state or a football club). While the context field generally contains freeform text, certain contexts occur repeatedly (e. g. *"Ortsname"* (place name) for city names). Such contexts can be used to assign supersenses (or other semantic categories) directly without having to consult *GermaNet*. Similarly, certain glosses have semantic prefixes that can be used directly, such as the $NAME- prefix for person names or $NUM- prefix for numbers and related terms. Finally, we introduce a pseudo-supersense called stopwords to which we assign signs whose German context word is found in a list of common stopwords.

Using these finetuning steps in concert with the pipeline described in Section 6.2, we are able to assign supersenses to 94% of *DGS Corpus* subtypes. Of these, 82% are assigned three supersenses or less and 46% are assigned a single one.

7. Outlook

In this article we presented approaches for creating collocation views for sign language research by using glosses and wordnet supersenses. In the future we hope to improve upon these in several ways. For example, up until now we only consider immediate neighbours of a target sign. However, collocates can also be separated from the target by other signs, so future collocation analyses should consider larger windows or skip-grams (cf. Järvelin et al., 2007). We also envisage dynamic merging of the right and left neighbour lists in cases where collocates seem to follow a free word order.

We hope to extend our cross-lingual bootstrapping of semantic information to finer semantic information than supersenses, such as near-synonyms or the hyperonymy hierarchy of *GermaNet*. This introduces new challenges, such as how to group terms within the hierarchy, and increases the relevance of known issues, like the overgeneralisation we face when selecting word senses.

Another exciting possibility is the potential of creating a feedback loop between the cross-lingual bootstrapping of wordnet information and the word sense discrimination performed by the lexicographers. While we showed how the bootstrapping can help lexicographers, we hope that in return the lexicographers' descriptions can improve the bootstrapping process by providing sign sense inventories and select token sense tags.

8. Acknowledgements

This publication has been produced in the context of the joint research funding of the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the Academies of Sciences and Humanities.

9. Bibliographical References

Atkins, B. T. S. and Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford University Press, New York, USA.

Auer, P. (2010). Zum Segmentierungsproblem in der Gesprochenen Sprache. Pre-Publication in *InLiSt – Interaction and Linguistic Structures*, Number 49.

Bartsch, S. (2004). *Structural and Functional Properties of Collocations in English: A Corpus Study of Lexical and Pragmatic Constraints on Lexical Co-Occurrence*. Doctoral thesis, Tübingen : Naar, Darmstadt, Germany.

Church, K. W. and Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1):22–29, March.

Fuertes-Olivera, P. A., Bergenholtz, H., Nielsen, S., and Niño Amo, M. (2012). Classification in Lexicography: The Concept of Collocation in the Accounting Dictionaries. *Lexicographica*, 28(1):293–308.

Hamaker, J., Zeng, Y., and Picone, J. (1998). Rules and Guidelines for Transcription and Segmentation of the SWITCHBOARD Large Vocabulary Conversational Speech Recognition Corpus. Technical report version 7.1, Institute for Signal and Information Processing, Mississippi State University, Starkville, Mississippi, USA.

Hamp, B. and Feldweg, H. (1997). GermaNet - a Lexical-Semantic Net for German. In *Proceedings of the ACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15, Madrid, Spain. ACL.

Hanke, T. and Storz, J. (2008). iLex – A Database Tool for Integrating Sign Language Corpus Linguistics and Sign Language Lexicography. In *Proceedings of SignLang@LREC*, pages 64–67, Marrakech, Morocco. ELRA.

Hanke, T. (2004). Hamnosys – Representing Sign Language Data in Language Resources and Language Processing Contexts. In *Proceedings of SignLang@LREC*, pages 1–6, Lisbon, Portugal. ELRA.

Järvelin, A., Järvelin, A., and Järvelin, K. (2007). s-grams: Defining generalized n-grams for information retrieval. *Information Processing & Management*, 43(4):1005–1019.

Johnston, T. (2010). From Archive to Corpus: Transcription and Annotation in the Creation of Signed Language Corpora. *International Journal of Corpus Linguistics*, 15(1):106–131.

Konrad, R., Hanke, T., Langer, G., König, S., König, L., Nishio, R., and Regen, A. (2018). Public DGS Corpus: Annotation Conventions. Project Note AP03-2018-01, DGS-Korpus project, IDGS, Hamburg University, Hamburg, Germany.

Landau, S. I. (2001). *Dictionaries: The Art and Craft of Lexicography*. Cambridge University Press, 2nd edition.

Langer, G., Troelsgård, T., Kristoffersen, J., Konrad, R., Hanke, T., and König, S. (2016). Designing a Lexical Database for a Combined Use of Corpus Annotation and Dictionary Editing. In *Proceedings of SignLang@LREC*, pages 143–152, Portorož, Slovenia. ELRA.

Langer, G., Müller, A., and Wähl, S. (2018). Queries and Views in iLex to Support Corpus-Based Lexicographic Work on German Sign Language (DGS). In *Proceedings of SignLang@LREC*, pages 107–114, Miyazaki, Japan. ELRA.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to WordNet: An Online Lexical Database. *International Journal of Lexicography*, 3:235–244.

Nishio, R., Hong, S.-E., König, S., Konrad, R., Langer, G., Hanke, T., and Rathmann, C. (2010). Elicitation Methods in the DGS (German Sign Language) Corpus Project. In *Proceedings of SignLang@LREC*, pages 178–185, Valletta, Malta. ELRA.

Orlandi, A. (2016). Monolingual collocation lexicography: State of art and new perspectives. In Adriana Orlandi et al., editors, *Defining collocation for lexicographic purposes – From linguistic theory to lexicographic practice*, number 219 in Linguistic Insights, pages 19–70. Peter Lang, Bern, Switzerland.

Rychlý, P. (2008). A Lexicographer-Friendly Association Score. In *Proceedings of RASLAN*, pages 6–9, Karlova Studánka, Czech Republic.

Schmidt, T. (2014). The Research and Teaching Corpus of Spoken German — folk. In *Proceedings of LREC*, pages 383–387, Reykjavik, Iceland. ELRA).

Schwager, W. and Zeshan, U. (2008). Word Classes in Sign Languages: Criteria and Classifications. *Studies in Language*, 32(3):509–545.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27(3):379–423.

Sinclair, J. M. (1996). The search for units of meaning. *Textus*, 9(1):75–106.

Slobin, D. I. (2008). Breaking the Molds: Signed Languages and the Nature of Human Language. *Sign Language Studies*, 8(2):114–130.

Westpfahl, S. and Gorisch, J. (2018). A Syntax-Based Scheme for the Annotation and Segmentation of German Spoken Language Interactions. In *Proceedings of LAW-MWE-CxG 2018@COLING*, pages 109–120, Santa Fe, New Mexico, USA. ACL.

Zwitserlood, I. E. P., Kristoffersen, J. H., Troelsgård, T., and Jackson, H. (2013). Issues in sign language lexicography. In *Bloomsbury Companion To Lexicography*, Bloomsbury Companions, pages 259–283. Bloomsbury Continuum, London, United Kingdom.