Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5276–5281
Marseille, 11–16 May 2020

Typical Sentences as a Resource for Valence

Uwe Quasthoff, Lars Hellan, Erik Körner, Thomas Eckart,
Dirk Goldhahn, Dorothee Beermann
Natural Language Processing Group, University of Leipzig, Germany
Department of Language and Literature, Norwegian University of Science and Technology, Norway
{quasthoff, koerner, teckart, dgoldhahn}@informatik.uni-leipzig.de
{lars.hellan, dorothee.beermann}@ntnu.no

Abstract

We describe a methodology by which verb valence information can be derived from corpora by using subcorpora of typical sentences that are constructed in a language independent manner based on frequent POS structures. The inspection of typical sentences with a fixed verb in a certain position can show the valence information directly. Using verb fingerprints, consisting of the most typical sentence patterns the verb appears in, we are able to identify standard valence patterns and compare them against a language's valence profile. With a very limited number of training data per language, valence information for other verbs can be derived as well. Based on the Norwegian valence patterns, we are able to find comparable patterns in German where typical sentences are able to express the same situation in an equivalent way, and can so allow for the construction of verb valence pairs for a bilingual dictionary. This contribution discusses this application with a focus on the Norwegian valence dictionary NorVal.

Keywords: valence, valence frames, typical sentences, minimal sentences, German, English, Norwegian, corpus

1. Introduction

Recent studies of sentence pattern frequencies indicate concentrations of POS-based signatures for short recurrent patterns reflecting realizations of patterns like ‘transitive’, ‘copular’ and sentence embedding, applied to corpora of German, English and Norwegian; sentences instantiating these patterns are called *typical sentences* in Müller et al. (2018). From a grammatical viewpoint, notions like ‘transitive’, ‘copula’ and embedded infinitive stand for valence patterns, as well as for typical POS-signatures. Given that, one can address the questions whether corpus search for typical POS-signatures might be used to identify corpus instantiations of interesting varieties of valence frames. This is not trivial, since in terms of morpho-syntactic parameters of verb valence bound items, languages like those mentioned have between 200 and 300 types of valence frames.

We here present an approach which in a corpus search allows one to ‘strip’ the POS-pattern of a sentence with various types of adjuncts down to those items representing the valence frame instantiated by the verb in question. These ‘stripped’ strings can be correlated with a set of minimal sentences representing the full range of valence frames of a language, and if a match is established between the ‘reduced’ POS-signature of the sentence in question and the POS signature of the minimal sentence representing a given valence frame VF, then the actual sentence can be hypothesized as instantiating the valence frame VF. This may allow for a procedure of semi-automatic construction of valence corpora for many languages, the prerequisites being only corpus texts for the language on the one side and enumerations of the valence frames used in the language – called the language’s valence profile – on the other.

Section 2 presents the notion ‘typical sentence’ described by sentence signatures as used in recent studies and preliminary views of frequencies of POS-signatures for the three languages, built on different corpora and different annotation systems. They clearly represent the ‘transitive’ and ‘copular’ nature of the most-frequent signatures. We here also illustrate the idea of ‘stripping away’ adjuncts from sentences containing adjuncts.

Section 3 describes a vector space presentation for a verb by the set of sentence typical signatures this word occurs in; this so-called *verb fingerprint* is used for the extraction of valence properties.

Section 4 presents the current scene of valence dictionaries and corpora, and, with a focus on German and Norwegian, describe how such resources can be built and used with ‘minimal sentences’ as a key component.

Section 5 then illustrates the approach applied to the creation of corpus annotation and dictionary representations of infinitival constructions in German and Norwegian. In doing so, both languages provide ‘gold standards’ for valence resources against which the general methodology can be checked.

2. “Typical sentences” in Norwegian, German and English

Large text corpora compiled from publicly available written sources, e.g. news, Wikipedia, crawled Web pages or literature contain sentences of varying syntactic complexity and tend to be biased towards long sentences and complex syntactic structures. Therefore, Müller et al. (2018) introduced the concept of “typical sentences”, defined as sentences with a common syntactic pattern. In this approach, sentences are represented by their corresponding POS tag sequence. Typical sequences are identified based on their respective frequency in a text corpus. As one would expect, the most frequent sentence signatures belong to relatively short sentences with typical structure. Usually, some large blocks of near duplicate sentences are included. Examples are sentences differing mainly by numerals, for instance “This story has been viewed 618 times.”. Such near duplicates are removed using the entropy for different positions in the sentence. The removal of these near duplicates and the selection of the (typically 100,000) most frequent sentence signatures gives a language independent extraction method for typical sentences. Based on the approach of Müller (2018), such typical sentences represent about 5%-10% of the original corpus.

Tables 1 - 3 show examples of high frequent sentence structures based on these POS tag patterns for all three languages addressed in this paper.[1] For better illustration concrete sentences extracted from the respective corpora are also provided.

Sentence signature	Sentences in the corpus	Example sentence (*English translation*)
{PRON VERB DET ADJ NOUN PUNCT}	4,412	Alle erkjenner de faktiske forhold. (*Everyone recognizes the facts*)
{PRON VERB ADJ NOUN PUNCT}	3,549	Alle blir gode fotballspillere. (*Everyone becomes good footballers.*)
{PRON VERB NOUN PUNCT}	2,626	Alle elsker gull! (*Everyone loves gold!*)
{NOUN VERB ADJ PUNCT}	1,686	Adkomst er vesentlig. (*Access is essential.*)

Table 1: High frequent sentence signatures in Norwegian (Bokmål, nob), based on news texts of 2015 published at Norwegian TLD.

Sentence signature	Sentences in the corpus	Example sentence (*English translation*)
{ART NN VAFIN ADJD $.}	42,220	Das Anliegen ist verständlich. (*The request is understandable.*)
{ART NN VAFIN ADV ADJD $.}	28,337	Das Angebot war sehr gut. (*The offer was very good.*)
{ART NN VVFIN CARD NN $.}	26,000	Das Buch kostet 20 Euro. (*The book costs 20 Euro.*)
{PPER VVFIN ART NN $.}	20,904	Er besorgt den Revolver. (*He obtains the revolver.*)

Table 2: High frequent sentence signatures in German, based on different sources acquired until 2018.

Sentence signature	Sentences in the corpus	Example sentence
{DT NN VBZ JJ SENT}	6,613	A baby is imminent.
{PP VVP DT NN SENT}	4,955	He changed the world.
{NP VVD DT NN SENT}	3,785	Aamir captioned the image.
{DT NN VHZ VBN VVN CD NNS SENT}	3,459	The band has been nominated four times.

Table 3: High frequent sentence signatures in English, based on news texts of 2016.

As expected, only a small part of these high frequent patterns describe a rather complex syntactic structure (like inclusion of relative clauses) or sentence structures that do not follow the standard word order in those languages (like SOV). But all of the frequently used sentence structures are included. Hence, we hope to find the different usage patterns in nearly minimal form.

For the verb fingerprint described in the following section, we will be interested in simple sentences containing a certain verb, specifically typical sentences with the same sentence structure which contain the verb in the same position.

Unfortunately, these additional restrictions reduce the number of occurrences of a given verb in a certain sentence structure too much. In many cases, sentence structures can be considered as identical (for the analysis of valency), if they differ slightly. Examples for a slight difference given by one additional word are

- An additional adverb (with a possibly necessary reordering of the words in the sentence after removal). Example: “Gestern habe ich Max getroffen.” (“*Yesterday I met Max.*”) will be reduced to “Ich habe Max getroffen.” (“*I met Max.*”).
- One additional adjective in an NP. Example: “Das Angebot war sehr gut."(“*The offer was very good.*”) will be reduced to “Das Angebot war gut.” (“*The offer was good.*”).

Much more complex rules are possible, but these two rules are the most important and can be applied iteratively. As shown in Schiffer et al. (2017), these two rules have turned out to be the most productive for German, and more complex rules will increase the number of reducible sentences by only less than 10%. The same could be expected for the other languages under consideration.

In comparison to the original corpus, a corpus of typical sentences exhibits certain differences. Among them are a shorter average sentence length and changes in the ranking of stopwords which are based on the relatively simple structure of the sentences. Examples for changes in stopword ranking are capitalized articles and pronouns which increase in frequency due to many sentences beginning with simple nominal phrases. On the other hand, conjunctions which usually form more complex and long sentences appear less frequent in typical sentences.

3. Verb fingerprint

For a structural description of a verb, the most frequent structures of typical sentences containing a certain verb, maybe in inflected form, are collected. Representing the IDs of the (say, 20 most frequent) sentence structures (together with normed frequency) in a vector gives a vector space representation of the structures a word appears in. Two verbs with similar representation will show up in many identical constructions. If the frequency of the two verbs is above some threshold, this clearly means similar valence profile (see next section). But these verb fingerprints show even more information:

- Is the verb often used in active and/or passive voice?
- Are object slots optional or not?

If one inspects the words filling a certain slot, information about the words can be found:

- Animate or inanimate nouns: What are typical pronouns *he*/*she* or *it*?

[1] Search terms as well as the exact choice of verb types here differ between the languages, but the figures nevertheless show the overall tendency for transitive and copular constructions to be the most frequent within sentences of this size.

- Semantic class for nouns: What are typical features of the nouns in a fixed position?

For example: From a set containing sentences like "He always drinks coffee." and "She never drinks alcohol." one can conclude that the subject of *drink* is animate and that coffee and alcohol are possible or even typical objects of *drink*, depending on their frequency.
Using these verb fingerprints it should be possible to identify standardized valency patterns.

4. Valence profiles

4.1 Relating typical sentences to valence

Described in terms of valence, the above examples, even when subclassified according to the verb types indicated, are essentially transitive and copular constructions. An interesting question is whether it would be possible - and make sense - to operate with a notion of typical sentences also relative to other valence types. A search as follows shows that among the 10 most frequently occurring verbs in a German corpus, no less than 5 are verbs that can or must be followed by finite or infinitival clauses[2]:

(1) *sagen* ('say')*, geben* ('give')*, kommen* ('come')*, gehen* ('go')*, stehen* ('stand')*, lassen* ('let')*, machen* ('make')*, bleiben* ('remain')*, liegen* ('lie')*, sehen* ('see')

This suggests that from a frequency perspective there may well be distinctive subclasses of frequent patterns that involve such more complex valence frames. To assess this, one needs to know how many valence types there generally are in a language, and what a minimal sentence instance of each type will look like - the latter reflecting an end point of a 'stripping' process as described above, relative to each type. This will be the topic of the following section, where we describe some aspects of valence in general and two kinds of resources involving valence, valence dictionaries and valence corpora. We here also show the usefulness of 'minimal sentences', which will constitute a further motivation for procedures for identifying typical sentences in the sense introduced above, but now instantiated for a much larger class of valence types. In section 5 we exemplify the idea relative to infinitival verbs in German and Norwegian.

4.2 The notion of valence and resources involving valence

A verb's *valence* is constituted by those types of expressions co-occurring with the verb which are necessary, either to express a complete instance of the type of situation expressed by the verb, or for formal reasons, or both. Such expressions are called *arguments* of the verb or items *valence-bound* by the verb; together they constitute the *valence frame* of the verb (or simply the valence). Most dictionaries or lexicons will provide some information about the valence of the verbs; a *valence dictionary* (or valence lexicon) will enter such information systematically for all the verbs[3]. From a normal user perspective, such dictionaries are useful in the way they systematize usages of a verb (most verbs can be used with more than one valence frame, and few verbs have exactly the same set of valence frames). From an NLP viewpoint, they can be crucial in structuring the results of syntactic parsing, and even more so when parsing is combined with the assignment of semantic representations; integration of digital dictionaries into, e.g., HPSG[4] and LFG[5] grammars is known as readily doable.
Common parameters of valence specification include number of arguments, their syntactic form, case, and special semantic relations. The way they are realized depends on the grammar of the language.
By a language's *valence profile* we understand the set of valence frame types available in a language. In defining the valence profile of a language, formal syntactic parameters could be summed up using simple phrase structure rules, whereas properties tied to function and semantic properties are less amenable to such formats, favoring feature structure specifications. Shorthand codes for the combination of such dimensions of specification are available in both descriptive and formal frameworks, and some will be exemplified below. In presenting a valence profile, a preferable format is an enumeration of valence types as a list of sentences each headed by a label for the valence type it instantiates. An example, with references to profiles for other languages, is given at:
https://typecraft.org/tc2wiki/Valence_Profile_English

With a valence dictionary, one is, from a research viewpoint, in a position to investigate the distribution of valence frames across verbs, to see what kinds of verb meanings tend to allow for many alternative frames and which these are in the various cases – this line of investigation may be called that of *lexical distribution of valence profiles*.
A valence dictionary should be aligned with a *corpus*, in which the verbs and their arguments can be identified relative to valence properties (possibly even annotated according to a chosen code of representation). Such a corpus will allow for the investigation of the *textual distribution* of the valence profile concerned, allowing one to see what valence frames are frequently realized, and together with chosen other parameters.
A *comparative* valence dictionary will align the valence information pertaining to two (or more) languages, comparing their respective valence profiles in general as well as numerical aspects of their lexical distribution (like 'does Norwegian have as many ditransitive verbs as German?'), but, more interestingly from many viewpoints, also comparing how given meanings are realized in terms of valence frames across the languages. If the dictionaries are accompanied by corpora, comparative aspects of textual distribution can also be pursued.
The present paper relates to the construction and comparison of valence dictionaries and corpora of German and Norwegian, with a view to discussing the roles that typical sentences can play relative to various aspects of these enterprises.[6] In our approach our main

[2] Based on the DWDS corpus: DWDS https://www.dwds.de/ Digitales Wörterbuch der Deutschen Sprache

[3] Among existing valence dictionaries are for instance: English:FrameNet; VerbNet; PropBank; German: Evalbu; Chech:Vallex; Polish: Walenty;
respective urls: https://framenet.icsi.berkeley.edu/fndrupal, http://verbs.colorado.edu/~mpalmer/projects/verbnet.html, https://ufal.mff.cuni.cz/czengvallex, http://hypermedia2.ids-mannheim.de/evalbu/, http://ucnk.ff.cuni.cz, http://clip.ipipan.waw.pl/Walenty.

[4] For HPSG ('Head-driven Phrase Structure Grammar), see Pollard and Sag (1994), Copestake (2002).

[5] For LFG (Lexical Functional Grammar) see Bresnan (2001).

[6] Among existing comparative valence dictionaries can be mentioned: CzEngVallex (https://ufal.mff.cuni.cz/czengvallex.), MultiVal (http://regdili.hf.ntnu.no:8081/multilanguage_valence_demo/mul

point of departure is the Norwegian valence dictionary NorVal, still under construction, and an accompanying valence corpus.[7] The German corresponding valence dictionary – GerVal - is to some extent modeled on NorVal.

In NorVal, entries are ordered not according to lemmas, but to *lemma+valence frame*, thus each lemma is represented in as many entries as there are valence frames for it. We call a *lemma+valence frame* a *lexval*. A lexval has the following basic structure, illustrated (for the occasion with italics and bold); on the left side is the lemma form together with possible selected items in the frame, and on the right side, divided by '__', the label for the frame, using the code ConstructionLabeling:[8]

(2)
Lemma-form + selected item
|
*befatte-med*__**trObl-obRefl-oblEqObInf &**
|
Frame label

A randomly selected snippet of the dictionary on this minimal form will look as follows, out of currently 12,884 lines:

(3)
befatte-med__trObl-obRefl-oblEqObInf &
befatte-med__trObl-obRefl-oblN &
befeste__tr &
befinne__trObl-obRefl-oblLoc &
befinne__trScpr-scObNrg-obRefl-scAdj &

A first instance of the role of minimal sentences is illustrated next. For each entry, a minimal sentence illustrates the use of the verb in the type of frame indicated (thereby also serving to illustrate the formula; the '&' is a comma in a possible CSV construal):[9]

(4)
befatte-med__trObl-obRefl-oblEqObInf & hun befatter seg med å kode &
befatte-med__trObl-obRefl-oblN & hun befatter seg med dem &
befeste__tr & du befester dem &
befinne__trObl-obRefl-oblLoc & hun befinner seg her &
befinne__trScpr-scObNrg-obRefl-scAdj & hun befinner seg vel &

Another possible role of minimal sentences is as follows: The comparative German-Norwegian dictionary – called 'PolyVal' – will construct pairs of entries from the respective dictionaries, constituted by minimal sentences instantiating the entries. These minimal sentences constitute *equivalences*, in the sense that they can express *the same situation* in a given communicative context. Examples of such equivalences are given in the table below:

	Norwegian	German
Equivalence 1	*hente*, tr Jeg henter boka *I fetch the book*	*holen*, tr Ich hole das Buch
Equivalence 2	*kjøpe*, tr Jeg kjøper bilen *I buy the car*	*kaufen*, tr Ich kaufe das Auto
Equivalence 3	*kjøpe*, tr Jeg kjøper avisa *I buy the newspaper*	*holen*, tr Ich hole die Zeitung
Equivalence 4	*stole på*, intrObl Jeg stoler på ham *I rely on him*	*vertrauen*, tr-obDat Ich vertraue ihm
Equivalence 5	*stole på*, intrObl Jeg stoler på Ola *I rely on Ola*	*sich verlassen auf*, trObl-obRefl Ich verlasse mich auf Ola

Table 4: German-Norwegian Equivalences

Each equivalence may count as representing a meaning – named in both languages, and distinguished from all other meanings by the fact that it constitutes a different line than all other meanings.

For the creation of minimal sentence equivalents we suggest using existing machine translation systems. At the moment we are using MyMemory (https://mymemory.translated.net) feeding the system 300 sentences at the time, using line breaks as separators.[10]

Thus, for each illustrated lexval in NorVal (like above) we use the sentence as input to the automatic translator, whereby two items are added at each line: a translation, and the lemma form of the verb used in the translation – the following illustration uses the same snippet as shown above (missing relative to the 'Equivalences' view in Table 4 is just the valence frame label for the German entry, which cannot be generated automatically):

(5)

['befatte-med__trObl-obRefl-oblEqObInf ', ' hun befatter seg med å kode ', ''] & Sie beschäftigt sich mit Codierung & ['beschäftigen']

['befatte-med__trObl-obRefl-oblN ', ' hun befatter seg med dem ', ''] & sie befasst sich mit ihnen & ['befasst']

['befeste__tr ', ' du befester dem ', ''] & Sie befestigen sie & ['befestigen']

['befinne__trObl-obRefl-oblLoc ', ' hun befinner seg her ', ''] & sie ist da & []

['befinne__trScpr-scObNrg-obRefl-scAdj ', ' hun befinner seg vel ', ''] & es geht ihr gut & ['gehen']

In addition it may be the case that two lexvals in one language correspond to one lexval in the other; an example of this is Equivalences 4 and 5 in Table 4.

For compactness one can summarize correspondences like in Table 4 on the following form:

tivalence), ContraGram (http://www.contragram.ugent.be/) and The Leipzig Valency Classes project (ValPaL; http://valpal.info/), of which only the first two have scope over entire verbal systems, and only CzEngVallex addresses meaning equivalence between the languages..

[7] For background see Hellan and Bruland 2015, Hellan et al. 2017.

[8] See Dakubu and Hellan 2017, and Construction Label tags

[9] Translations of the respective minimal sentences:
She deals with coding; she deals with them; you consolidate them; she is here; she is well.

[10] Cooperation with Melanie Siegel, Darmstadt.

(6)

Equivalence 1	like, tr
Equivalence 2	like, tr
Equivalence 3	like, tr
Equivalence 4	unlike, <intrObl, tr-obDat>
Equivalence 5	unlike, <intrObl, trObl-obRefl>

Given a large number of such lines, one can get a picture of which frames tend to be constant across the languages for given meanings and which ones not. This in turn may shed light on the more general question of whether valence frames are to some extent ‘predictable’ from meaning. This illustrates a third perspective in which minimal sentences may be of interest.

Given these many respects in which minimal sentences may play a key role in constructing and maintaining valence resources, it will be interesting to see if the notion and operations of typical sentences can be aligned with any of the uses of minimal sentences. Foremost may be the question whether typical/minimal sentences may be of help in identifying or detecting valence types in a corpus.

Given at the outset, as mentioned, that the first view of typical sentences highlight transitive and copula constructions, a conceivable ‘backwards’ check of the POS-signatures associated with transitivity may be to look at 4000 Norwegian verb lemmas that occur only transitively, see their actual POS-patterns in a corpus and assess their strippability.

Then, opening an investigation of more complex valence patterns, we exemplify the issue for some infinitival constructions in German and Norwegian.

5. Infinitival constructions

For German[11], GerVal contains 104 verbs which either exclusively select an infinitival complement or select an infinitive as one of their frames, altogether representing 136 frames. Not considered are constructions featuring adverbial infinitival phrases or infinitival phrases in predicative constructions, such as *Zu tanzen ist Spaß* (‘to dance is fun’) and *Bayern ist zu schlagen* (Bayern is to be beaten’).

Relative to this resource we have observed 5 dominant POS structures which we list in Table 5, starting from the most integrated VV-pattern for example projected by sensory verbs (e.g. *hören*, *sehen*), modal verbs and the causative verb *lassen*, as well as the deictic verbs *kommen* and *gehen*. We only represent the head verb (e.g. *müssen, scheinen, hören*, etc) and its arguments linearized in a SVO linear pattern. Pattern 5 in Table 5 where the infinitive is introduced by a preposition is projected by so-called attitude verbs or verbs of communication such as *reden*, *schreiben* and *nachdenken* with the preposition *über* or *von*, as in *Er redet darüber Ski fahren zu gehen.* ’Han snakker om å gå på ski’. In Table 5 we list the valence patterns together with a German and a Norwegian example ; the latter is preceded by an exclamation mark when their valence frame differs from the German frame.

Valence Pattern in terms of STTS-POS-Pattern[12]	German example	Norwegian example using the same verb
<s> w1:[[tag="VVFIN"] w2: [tag="VVINF"] < s/>	Er muss gehen	Han må gå
<s> w1:[[tag="VVFIN"] w2: [tag="VVIZU"] < s/>	Er scheint zu gehen	Han synes å gå
<s> w1:[[tag="VVFIN"] w2:[tag=("N*"| "PPER")] w3:[tag="VVINF"] < s/>	Er hört ihn gehen	Han hører ham gå
<s> w1:[[tag="VVFIN"] w2:tag="("N*"| "PPER")] w3 + w4:[@zu $p=VVINF] < s/>	Er bittet den Trainer zu gehen Er verspricht ihm zu gehen	!Han ber treneren gå Han lover ham å gå
<s> w1:[[tag="VVFIN"] w2:[lemma="dar*"] w3 +w4: [@zu $p=VVINF] < s/>	Er redet davon zu verreisen	Han snakker om å reise.

Table 5: Infinitival Patterns in German

In German one of the factors that need to be taken into consideration is that *zu* infinitives may be expressed as a morphological pattern rather than as a syntactic one; *Er versuchte diese Frage* ***an-zu-sprechen***. ‘Han prøvde å stille dette spørsmålet.’ A further point to mention is that the STTS tagset allows us to distinguish between embedded copula constructions and embedded main verbs, as shown in Table 6 below.

Sentence Signature	Example Sentence (English translation)
{PPER} **scheint** {ART}{ADJA} {NN} **zu** {VAINF} {$.}	Es scheint eine erfolgreiche Strategie zu sein. (*It seems to be a successful strategy.*)
{PPER} **scheint** {PIAT}{NN} **zu** {VVINF}{$.}	Es scheint keinen Ausweg zu geben (*There seems to be no way out*)
{PPER} **scheint** {PPER} **zu** {VVINF}{$.}	Es scheint ihn zu amüsieren. (*It seems to amuse him.*)

Table 6: Example signatures and sentences for German “scheinen zu” (*seem to*)

[11] Linguistic studies leading up to this part of the resource include Beermann 2017 and Beermann and Hellan 2016.

[12] http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/stts.asc and as PDF: http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf

The signature here sought for can be accommodated within a larger context such as

(7)
w-1 ...w-n+1, VVFIN, w-1 ...w-n+1, VVIZU w-1 ...w-n+1.

The flexibility and versatility in combinations of predominantly structural POS patterns and lexical specifications suggests that the strategies here outlined may well be able to accommodate also the more complex valence frames, which correpond in many cases also longer minimal sentences and thus also allow for more variation and larger sets of POS structures.

6. Conclusion

The notion 'typical sentence' was first introduced as a notion of equivalence class of very short sentences in large corpora, constituting altogether around 10% of a corpus. It was noticed, as illustrated in section 2, that they tend to instantiate sentences with copula or transitive verbs. Retaining and extending the notion, we use 'minimal sentence' as a notion systematically relativized to valence frames, in the sense that for each type of valence frame in a language, one can, for each verb carrying that type of frame, envisage a typical or minimal sentence pattern instantiating it. A corpus displaying instantiations of the totality of valence frame types is a desirable resource, however, as is well known, the annotation 'by hand' of a corpus with regard to the valence frame types instantiated in it, and specified in terms of standard valence type notions, is a laborious task. It is so far only for 'deep' grammatical parsers with lexical valence information that such assignments can be done automatically. What we propose in this paper is a methodology for, for any language for which a valence dictionary is available, identifying minimal sentence patterns representing its respective frame types, and representing these patterns in terms of POS sequences. Such sequences we call the 'fingerprints' of the various frame patterns, and derivatively, of the verbs carrying the frames in question (cf. section 3). Given the efficiency with which POS assignment can be applied to a given corpus, we thereby open for a procedure by which valence assignment can be accessible by virtue of the corpus having undergone POS assignment, and presupposing access to verb valence frames. As shown in section 5, moreover, even for the putatively most complex types of constructions in the languages – infinitival constructions – POS patterns can be defined characterizing different types of valence frames.

7. Bibliographical References

Beermann, D. (2017). Infinitives: a comparative German-Norwegian study. In Hellan, Malchukov and Cennamo (eds).

Beermann, D. and Hellan, L. (2016). Switched Control and other 'uncontrolled' cases of obligatory control. In: Doug Arnold, Miriam Butt, Berthold Crysmann, Tracy Holloway King, Stefan Müller (eds) Proceedings of the Joint 2016 Conference on Head-driven Phrase Structure Grammar and Lexical Functional Grammar. Stanford: CSLI Publications.

Beermann, D. and Mihaylov, P. (2014). Collaborative databasing and Resource sharing for Linguists. Languages Resources and Evaluation 48. Dordrecht: Springer, 1-23.

Bresnan, J. (2001). *Lexical Functional Grammar*. Oxford: Blackwell.

Copestake, A. (2002). *Implementing Typed Feature Structure Grammars*. CSLI Publications.

Dakubu, M.E.K. and Hellan, L. (2017). A labeling system for valency: linguistic coverage and applications. In: Hellan, Malchukov and Cennamo (eds).

Hellan, L. and Bruland, T. (2015). A cluster of applications around a Deep Grammar. In: Vetulani et al. (eds) Proceedings from The Language & Technology Conference (LTC) 2015, Poznan.

Hellan, L., Beermann, D., Bruland, T., Haugland, T., and Aamot, E. (2017). Creating a Norwegian valence corpus from a deep grammar. In: Vetulani et al. (eds) Proceedings from The Language & Technology Conference (LTC) 2017, Poznan.

Hellan, L., Malchukov, A., and Cennamo, M. (eds.) (2017). Contrastive Studies in Verbal Valency. John Benjamins Publ. Co. Amsterdam.

Müller, L., Quasthoff, U., and Sumalvico, M. (2018). Corpora of Typical Sentences. In: LREC 2018, Eleventh International Conference on Language Resources and Evaluation, Miyazaki, Japan, 2018.

Pollard, C. and Sag, I. (1994). *Head-driven Phrase Structure Grammar*. Chicago Univ. Press.

Schiffer, L., Quasthoff, U., and Müller, L. (2017). Syntactic Sentence Simplification and Sentence Compression for German, Proceedings of the LTC'17: 8th Language & Technology Conference, November 17-19, 2017, Poznań, Poland.

8. Language Resource References

Typecraft:
https://typecraft.org

Valence Profile Norwegian:
https://typecraft.org/tc2wiki/Valence_Profile_Norwegian

Valence Profile English:
https://typecraft.org/tc2wiki/Valence_Profile_English

Norwegian Valency Corpus:
https://typecraft.org/tc2wiki/Norwegian_Valency_Corpus

The Norwegian valence dictionary NorVal is under development, a pointer to its stages of development being given at the link above.