

Evaluating a German Sketch Grammar: A Case Study on Noun Phrase Case

Kremena Ivanova*, Ulrich Heid*, Sabine Schulte im Walde*, Adam Kilgarriff[◦], Jan Pomikálek^{◦▷}

*Institute for Natural Language Processing, University of Stuttgart, Germany

[◦]Lexical Computing Ltd, Brighton, UK

[▷]Masaryk University, Brno, Czech Republic

{ivanovka,heid,schulte}@ims.uni-stuttgart.de,
adam@lexmasterclass.com, xpomikal@fi.muni.cz

Abstract

Word sketches are part of the Sketch Engine corpus query system. They represent automatic, corpus-derived summaries of the words' grammatical and collocational behaviour. Besides the corpus itself, word sketches require a *sketch grammar*, a regular expression-based shallow grammar over the part-of-speech tags, to extract evidence for the properties of the targeted words from the corpus. The paper presents a sketch grammar for German, a language which is not strictly configurational and which shows a considerable amount of case syncretism, and evaluates its accuracy, which has not been done for other sketch grammars. The evaluation focuses on NP case as a crucial part of the German grammar. We present various versions of NP definitions, so demonstrating the influence of grammar detail on precision and recall.

1. Introduction

The Sketch Engine (Kilgarriff et al. (2004)¹) is a corpus query tool which supports the extraction of *word sketches*, corpus-based summaries of a word's grammatical and collocational behaviour, from large corpora. Word sketches serve as a starting point for the description of lexicographic properties at the syntax-semantics interface such as subcategorisation, selectional preferences and collocations. The Sketch Engine has been used, for lexicography and for linguistic research, for around twenty languages to date.

In this paper, we apply the Sketch Engine to German. German occupies an intermediate position between configurational languages like English, which encode grammatical relations (e.g. subject-hood, object-hood, etc.) through the position of constituents, and case languages where morphological marking identifies grammatical relations, leaving a freer constituent order. German also has a number of sentence patterns which specify the position of the finite verb, and once the verb position is specified, noun phrases are subject to a considerable amount of scrambling. In addition, around 80% of noun phrases are case-ambiguous (Evert, 2004). These factors present challenges in the development of word sketches for German.

We evaluate different options for designing a German sketch grammar. The evaluation relies on a gold standard corpus of 1,000 test sentences annotated with noun phrase structure and case. For German, NP case is central to creating useful word sketches. The gold standard corpus lets us compute precision and recall for NP case identification. Our results are comparable to other shallow grammars of German, but inferior to results achieved using more sophisticated parsing frameworks. We outline how we plan to improve German word sketches using a richer framework.

Like other language technologies, a sketch grammar and the resulting word sketches might be evaluated from either of two perspectives: the developers' or the users'. The developer is interested in improving the system, so the evaluation needs to distinguish the contributions of different modules and show where changes in the method improve performance. By contrast the user is interested in the system as a whole, and whether it can support them in their work. This paper takes the developer's view; a complementary evaluation which assesses word sketches as a tool for users (who are lexicographers), across five languages, is currently in progress.

2. The Sketch Engine

The Sketch Engine, like other corpus tools, can be used for concordancing, and has a full range of functions for specifying, sorting and extracting keywords from concordances. Its distinctive feature is however word sketches. Word sketches are summaries of a word's behaviour based on a *sketch grammar*, which is used to identify collocations in a range of grammatical relations to the headword. Given a sketch grammar and a part-of-speech-tagged corpus, the system extracts lexical relations, between e.g. nouns and their attributive adjectives, verbs and their subjects, verbs and their objects, elements of a conjunction etc. The word sketch for a word presents a list of its collocates,² organised according to the grammatical relation they stand in to the headword and sorted according to co-occurrence significance.³ For an example see Tables 1 and 2.

²By *collocation* we mean an expression comprising two words which tend to go together. We call these words the headword or node word and the *collocate*.

³The default statistic is based on the Dice co-efficient, following Curran (2004), cf. "Statistics used in the Sketch Engine" at <http://trac.sketchengine.co.uk/wiki/SkE/DocsIndex>.

¹<http://www.sketchengine.co.uk>

The sketch grammar defines linear patterns in the form of regular expressions, and the participants of the grammatical relations are marked. Context may be provided by describing the intervening and surrounding words and/or structures within the regular expression.

So far, sketch grammars have been produced for a range of languages, both configurational (for example English (Kilgarriff et al., 2004) and Chinese (Kilgarriff et al., 2005)) and languages which have a detailed case morphology (including Slovene (Krek and Kilgarriff, 2006) and Czech (Kilgarriff et al., 2004)).

3. Characteristics of German

The writer of a German sketch grammar faces three types of problems arising from the sentence structure of the language. Firstly, more work is to be done than for a configurational language, as German has several models of constituent order (verb-initial, verb-second and verb-final). Secondly, these models allow for considerable freedom in the placement of noun phrases and prepositional phrases (in the *Mittelfeld*). Thirdly, only about 21% of all NP tokens in a news corpus are unambiguous with respect to case (Evert, 2004). Another 21% are fully ambiguous, and 58% are two- or three-way ambiguous. The ambiguity rate of individual elements of NPs, i.e. determiners, adjectives and nouns, taken in isolation, is even higher.

The following example, taken from a German administrative text, illustrates these problems.

- (1) *wenn die Mitgliedsstaaten der Gemeinschaft solche Vorschriften erlassen, ...*
'if the member states of the Community enact such regulations, ...'

In (1), *die Mitgliedsstaaten* and *solche Vorschriften* can both be either nominative or accusative, thus subject or object. The NP *der Gemeinschaft* can be a genitive (adjunct to the preceding noun) or a dative (complement or free dative), and the verb *erlassen* can take two arguments (subject and object) or three (subject, object and indirect object in the dative). The order of the three NPs could be scrambled (at least under certain information structural conditions), and PPs and/or adverbs (e.g. local or temporal adjuncts) could be placed between the NPs. The same is possible under the two other verb placement models.

4. Corpora, tools and tagsets

For work on large German corpora, standard low-level annotation tools are available for tokenising, part-of-speech-tagging and lemmatisation. Our corpus is DeWaC, a 1.6 billion word corpus of German drawn from the web (Baroni and Kilgarriff (2006)).⁴ The corpus is tokenised, lemmatised and part-of-speech-tagged by the Tree Tagger (Schmid, 1994) using the Stuttgart-Tübingen TagSet (STTS, Schiller et al. (1999)).

STTS is coarse-grained. It provides just one tag each for common nouns and articles, whatever their case, gender or

number, and the only distinction it makes for adjectives is between attributive and predicative. For verbs it does not identify tense, person or number. As a point of comparison, all tagsets for English, however coarse-grained, distinguish singular and plural nouns. A tagset following the MULTTEXT model would provide slots for case, gender and number for nouns, adjectives and articles, and for number, tense and person for verbs (Multext, 1995). There are, however, various arguments for making more, or less, fine-grained tagsets; in addition, it is more difficult to determine nominal case, gender and number on a word basis (without disambiguating context) in German than in English (cf. Section 3). Thus, STTS allows POS-taggers to avoid guessing in difficult cases where they would be likely to make many errors. Concerning the sketch grammar, however, this means that we need to establish nominal case, if verb+subject and verb+object pairs are to be found with any confidence. A considerable part of the work invested in the creation of a German sketch grammar was devoted to providing or approximating exactly this type of data.

5. A sketch grammar for German

A simple example of a sketch grammar rule is the following, for extracting adjective-noun collocations:

2 : [tag="ADJA"] 1 : [tag="NE|NN"]

The rule has two anchors, a noun (marked by '1', with a disjunction of part-of-speech tags 'NE' for proper names and 'NN' for common nouns), and an attributive adjective (marked by '2', with part-of-speech tag 'ADJA'). The rule can be used to cover both the modifier and the modifiee relationship between adjective and noun. Thus, one obtains a list of modified nouns for an adjective within the adjective's word sketch, and a list of modifying adjectives for a noun within the noun's word sketch, as Tables 1 and 2 illustrate for the adjective *klein* 'small' and the noun *Dorf* 'village'. The first column in each table lists the collocate and an English gloss, the second gives the collocation frequency and the third, the salience score for the collocation. The tables show the 10 most prominent collocates, sorted by significance.

The grammar defines twelve grammatical relations, covering attributive and predicative adjectives, noun phrase functions (subjects and direct and indirect objects of verbs, genitive modification), prepositional phrases, conjunctions and disjunctions, and verb particles. The German grammar does not include sentential complements, and is restricted to active clauses only. Passives, which pose the same problems of scrambling and add sub-patterns for the placement of auxiliaries and participles (Heid and Weller, 2008) have not yet been addressed.

Most of the grammar rules are as simple as the above example suggests. Verb subcategorisation is not. It relies heavily on identifying noun phrases and noun phrase case. Relations between verbs and subcategorised-for nouns are central to any account of collocation, so the definition of noun phrases and their combination within subcategorisation is a crucial task for a sketch grammar. Once noun phrases (and their cases) are identified, assigning syntactic functions is relatively straightforward. Providing subcategorisation data has two sub-tasks, (1) the identification of noun

⁴At time of writing the computation of word sketches for the full 1.6b words is not yet complete. Word sketches in this paper are based on a 100m word subset.

phrases and their case; (2) the identification of clause patterns. We explored different strategies for each.

Modified nouns		Freq	Sign
<i>Ausschnitt</i>	‘extract’	188	37.49
<i>Junge</i>	‘boy’	325	33.91
<i>Dorf</i>	‘village’	274	32.80
<i>Meerjungfrau</i>	‘mermaid’	46	31.19
<i>Mädchen</i>	‘girl’	352	30.88
<i>Gruppe</i>	‘group’	627	29.34
<i>Nenner</i>	‘denominator’	61	27.93
<i>Detail</i>	‘detail’	169	27.76
<i>Würfel</i>	‘dice’	68	27.47
<i>Schönheitsfehler</i>	‘flaw’	23	27.22

Table 1: Adjective-noun example word sketch:
Nouns modified by adjective *klein* ‘small’.

Modifying adjectives		Freq	Sign
<i>klein</i>	‘small’	274	37.68
<i>umliegend</i>	‘surrounding’	39	37.30
<i>malerisch</i>	‘picturesque’	20	28.96
<i>entlegen</i>	‘remote’	16	28.58
<i>verschlafen</i>	‘sleepy’	12	26.18
<i>gelegen</i>	‘situated’	26	26.05
<i>zerstört</i>	‘destroyed’	17	25.52
<i>ganz</i>	‘whole’	118	25.44
<i>abgelegen</i>	‘remote’	13	25.15
<i>kurdisch</i>	‘Kurdish’	16	24.54

Table 2: Adjective-noun example word sketches:
Adjectives modifying the noun *Dorf* ‘village’.

5.1. Noun phrases

Simple German NP structures can be extracted by the following regular expression:

DET? ADV* ADJA* NOUN

covering a linear sequence of a (possible) determiner, zero to several adverbs, zero to several attributive adjectives, and a noun. Problems arise as soon as one is interested in the cases and grammatical functions of the NPs. Complete disambiguation of NP case with both recall and precision of 100% may not be possible, but there are various strategies that provide linguistic knowledge in addition to the lemmas and the part-of-speech tags in the corpus data. In general, additional knowledge constrains the search: we get higher recall and less precision with underspecified queries, and higher precision but less recall for the more specific queries. For testing purposes, we use different versions of the sketch grammar, where we include different amounts of linguistic knowledge to identify the case (and thus the grammatical function) of the NPs:

Morphological restrictions

inflections The inflection of determiners and adjectives, by itself or together with information about the number and gender of nouns (see next item) disambiguates many NPs; we write a disjunctive set of specific rules to account for these inter-relationships. For example,

specifying a determiner ending of *-em* in combination with an adjective ending of *-en* disambiguates the NP case to *dative*.

affix-gender There are regularities between derivational affixes and the gender of nouns; we partition derived nominals into three subsets according to gender. The suffixes *-heit*, *-schaft*, *-ine* amongst others indicate feminine nouns and the suffixes *-ismus*, *-ist* indicate masculine nouns. In combination with determiner and/or adjective endings, the gender information disambiguates case. For example, the basic pattern DET NOUN with the determiner *den* (which is ambiguous between accusative singular and dative plural) is disambiguated to dative plural where the noun is feminine.

Structural restrictions

no-structure We do not take any sentence structure into account at all.

verb-final We only consider verb-final clauses. This puts the strongest restrictions on sentence structure. German verb-final clauses are sub-clauses that constitute around 20% of the data. They contain all subcategorisation-relevant material between a subordinating conjunction and the finite verb form and thus provide a clearly delimited domain for finding the complements (Eckle, 1999). Furthermore, NPs put restrictions on each other: a sub-clause with transitive frame might contain a nominative and an accusative NP in either order, but not two nominative or two accusative NPs. The sketch grammar for verb-final clauses uses the NP models and checks for patterns where one, two or three NP complements appear with the verb, in all possible constituent order permutations. In addition to the complements, we allow various types of modifications (such as adverbial phrases).

all-clauses Here we also take structural information into account, but not restricted to verb-final clauses. We model all clause types (verb-initial, verb-second, verb-final) at the same time. This knowledge type puts less restrictions on the clause types that are taken into account but at the same time allows more confusion between NP cases, as we are going beyond the clearly delimited domain of verb-final clauses.

We consider *inflection* to be the minimum information that we might use. *affix-gender* disambiguates some previously unclear noun cases but substantially restricts recall. The same applies to the structural restrictions in *verb-final* and *all-clauses*. We tested various combinations of the above linguistic knowledge levels, cf. Table 3. Note that we did not use a lexicon to establish gender or case in any of the conditions; we plan to do this in future work.

no affix-gender		no-structure;
	×	verb-final;
with affix-gender		all-clauses

Table 3: Combinations of linguistic knowledge.

5.2. Processing inside and outside the Sketch Engine

Additional linguistic input may be provided in corpus pre-processing by a POS-tagger or parser, or it may be provided within a sketch grammar. We began by preparing sketch grammar rules for identifying case (and, as part of that process, sometimes number and gender) for noun phrases. These rules became complex, and began taking a long time to run. On investigating the problem, we realised that we were sometimes re-computing a noun's case many times over, as the system tried to find all the ways in which each sketch-grammar clause might match a sentence. So we took the rules for assigning case and applied them as a pre-process, with the result being stored as an attribute of the word. The attribute was then indexed and was available for using in higher-level processing to find grammatical relations. The functionality was unchanged, but the implementation was markedly more efficient and it was possible to build word sketches for large corpora in reasonable time. Viewed in this way, we implemented a specialist POS-tagger which takes STTS-tagged data as input and returns data tagged according to a finer-grained tagset, in particular with nouns marked for case.

5.3. Gold standard corpus

As the gold standard, we use 1,000 randomly selected sentences from the DeWaC corpus, manually annotated for NPs by one of the co-authors. We annotated start and end point, and case. Examples (2)-(3) present three sentences from our gold standard annotation. The beginning and end point are marked by the brackets; the end bracket is accompanied by the NP case label. In total, the gold standard contains 1,709 NPs with nominative case, 618 NPs with accusative case, 149 NPs with dative case, and 437 NPs with genitive case.

- (2) *Doch auch [den Terroristen]_{NPdat} gelingt [die Flucht]_{NPnom}.*
'But also the terrorists succeed in the getaway.'
- (3) *[Ich]_{NPnom} musste [meine Arbeit]_{NPakk} schon sehr gut machen, um anerkannt zu werden.*
'I have to do my work really well to be approved.'

The sketch grammar is applied to this test corpus in six conditions, resulting from the combinations in Table 3:

1. inflection + no-structure
2. inflection + affix-gender + no-structure
3. inflection + verb-final
4. inflection + affix-gender + verb-final
5. inflection + all-clauses
6. inflection + affix-gender + all-clauses

5.4. Results

Tables 4 and 5 present two example word sketches, for the verb *öffnen* 'open' and the noun *Pflanze* 'plant', in German only. The word sketches list all grammatical relations the respective headwords appear with in our corpus, followed by the 20 most significant collocates that have a minimum joint frequency of 3 with the headword. *subj/subj-of* refers to the subjects of the verb, *obj-acc* and *obj-dat* to accusative and dative objects, *adv* to adverbs, *attr-adj*

to attributive adjectives, *gen-atr/gen-atr-of* to genitive attributes, and *and/or* to conjunctions and disjunctions. The sketches were created by condition 6. Table 4 illustrates the distinct NP cases that underlie *subj vs. obj-acc vs. obj-dat*: The strong overlap between *subj* and *obj-acc* on the one hand results from the verb *öffnen* that can but need not be used as a reflexive verb; on the other hand, this shows the difficulty to distinguish nominative and accusative case. *obj-dat* is dominated by personal pronouns, as *öffnen* is often accompanied by a benefactive.

Table 6 presents the precision and recall in the six conditions. An NP that is extracted by our grammar is counted as "correct" if the end point of the NP (indicating the nominal head) and the case label are both identical to the gold standard annotation. The left half of the table presents the results for the three conditions without affix-gender and the right half presents the results for conditions with it. Since our sketch grammar does not incorporate the subcategorisation of genitive NPs (as there are only few German verbs that subcategorise for a genitive case, we only consider NPs for genitive modification), conditions 3 and 4 that restrict the grammar to verb-final clauses are not implemented for genitive case and thus missing in the table.

Mistakes that are made by the grammar might be contributed either to the corpus data (which includes incomplete and incorrect sentences), to the preprocessing (which can go wrong in assigning part-of-speech tags, or lemmas), or to the sketch grammar itself. It would be a further task to separate these various sources of noise, and thus the results refer to overall success.

5.5. Comparing the conditions

Comparing condition 1 with condition 2 describes the effect of adding derivational gender information to the NP definition. Recall falls for all NP cases; precision increases.

Comparing condition 1 with condition 3 and comparing condition 2 with condition 4 describes the effect of restricting the NP search to verb-final clauses. Recall goes down greatly (as we are only considering ca. 20% of all clauses now) and precision rises. The decrease of recall and the increase of precision are both stronger in condition 3 than in condition 2, showing that the limitation by verb-final clause structure is more severe than the limitation by affix gender. Comparing condition 1 with condition 5 and comparing condition 2 with condition 6 describes the effect of adding sentence structure to the NP descriptions without restricting the NP search to a certain clause type. Again, recall falls and precision rises. In comparison to conditions 3 and 4, the effect is less strong, because we allowed all clause types. In contrast to our expectations, the precision values for accusative and dative case are better in conditions 5/6 than in conditions 3/4.

The overall best precision and recall results are in bold in the table. Unsurprisingly, the recall is largest in condition 1, with least restrictions. The best precision, however, is, in three out of four cases, not achieved by condition 4, the most restricted version of the grammar, but by condition 6, which takes all clause types into account. This result demonstrates that structural information helps the grammar but need not be restricted to the clearly delimited type.

subj	3017	5.1	obj-acc	282	5.9	obj-dat	136	5.4	adv	140	5.2
<i>Tür</i>	238	49.37	<i>Tür</i>	39	36.24	<i>ihr</i>	13	19.98	<i>täglich</i>	12	22.68
<i>Pforte</i>	35	35.20	<i>Auge</i>	26	26.67	<i>sie</i>	8	19.40	<i>versehentlich</i>	3	16.92
<i>Türe</i>	29	33.78	<i>Pforte</i>	7	22.71	<i>er</i>	9	18.96	<i>leicht</i>	6	13.89
<i>Tor</i>	62	32.34	<i>Wohnungstür</i>	3	21.61	<i>wir</i>	16	16.62	<i>weit</i>	13	13.61
<i>Auge</i>	114	32.29	<i>Türe</i>	5	19.38	<i>Markt</i>	3	9.04	<i>gleichzeitig</i>	4	12.37
<i>Fenster</i>	49	28.69	<i>Datei</i>	4	12.23	<i>ich</i>	6	7.86	<i>automatisch</i>	3	11.42
<i>Schleuse</i>	10	23.27	<i>Tor</i>	4	11.7	<i>endlich</i>	3	11.25			
<i>GAT-Bereich</i>	4	19.16	<i>Fenster</i>	3	9.32	<i>langsam</i>	3	10.97			
<i>Haustür</i>	8	18.88	<i>Herz</i>	3	7.72	<i>plötzlich</i>	3	10.94			
<i>Klappe</i>	8	18.20				<i>erneut</i>	3	9.90			
<i>Hangartor</i>	3	15.93				<i>schnell</i>	3	8.30			
<i>Datei</i>	13	15.51				<i>erst</i>	3	4.05			
<i>Schere</i>	6	14.77	and/or	90	0.7						
<i>Luke</i>	4	14.49	<i>schließen</i>	33	35.85						
<i>Herz</i>	19	14.41	<i>herunterladen</i>	3	18.83						
<i>Schublade</i>	6	14.08	<i>lesen</i>	4	12.65						
<i>Zimmertür</i>	3	14.00									
<i>Wurmloch</i>	3	13.91									
<i>Holzstür</i>	3	13.57									
<i>Reiseland</i>	4	13.42									

Table 4: Word sketch for verb *öffnen* ‘open’.

attr-adj	1566	2.0	subj-of	905	2.5	and/or	379	2.7	gen-atr-of	601	2.0
<i>gentechnisch</i>	94	47.14	<i>wachsen</i>	26	24.45	<i>Tier</i>	218	56.82	<i>Anbau</i>	20	30.63
<i>verändert</i>	100	42.3	<i>gedeihen</i>	6	18.46	<i>Pflanzenteil</i>	9	29.08	<i>Lebensraum</i>	10	20.25
<i>genmanipuliert</i>	30	39.44	<i>anbauen</i>	5	18.30	<i>Baum</i>	12	20.87	<i>Bestäubung</i>	4	20.17
<i>fleischfressend</i>	16	35.93	<i>werden</i>	73	15.91	<i>Mikroorganismus</i>	6	20.13	<i>Blatt</i>	13	20.15
<i>transgenen</i>	16	34.59	<i>können</i>	44	15.15	<i>Tierwelt</i>	4	17.24	<i>Wachstum</i>	13	19.51
<i>exotisch</i>	24	30.00	<i>sollen</i>	30	15.03	<i>Blume</i>	6	16.92	<i>Wurzelbereich</i>	3	19.16
<i>transgener</i>	8	28.45	<i>gießen</i>	4	14.52	<i>Frucht</i>	6	14.28	<i>Metamorphose</i>	5	18.09
<i>giftig</i>	18	26.44	<i>graben</i>	4	14.37	<i>Strauch</i>	3	13.10	<i>Nährstoffbedarf</i>	3	17.90
<i>heimisch</i>	20	23.37	<i>sein</i>	109	14.01	<i>Tierart</i>	3	12.02	<i>Same</i>	5	15.79
<i>manipuliert</i>	10	23.30	<i>blühen</i>	4	13.48	<i>Stein</i>	4	11.04	<i>Wurzel</i>	8	15.67
<i>abgestorben</i>	9	23.27	<i>müssen</i>	22	12.65	<i>Gebäude</i>	3	6.91	<i>Freisetzung</i>	4	14.96
<i>wachsend</i>	25	22.89	<i>fressen</i>	4	12.15	<i>Produkt</i>	3	6.02	<i>Gedeihen</i>	3	14.43
<i>genverändert</i>	6	22.73	<i>vermehrten</i>	3	11.36				<i>Biochemie</i>	3	14.21
<i>geschädigt</i>	10	21.23	<i>brauchen</i>	9	11.27	gen-atr	108	0.4	<i>Wasserversorgung</i>	4	14.03
<i>krautige</i>	4	20.69	<i>ausbreiten</i>	3	11.15	<i>Bibel</i>	4	14.83	<i>Monitoring</i>	3	13.95
<i>selten</i>	20	20.58	<i>befallen</i>	3	10.92	<i>Art</i>	9	14.56	<i>Aussterben</i>	3	13.61
<i>schnellwachsend</i>	4	19.39	<i>hervorbringen</i>	3	10.86	<i>Monat</i>	5	13.32	<i>Pflege</i>	6	12.60
<i>robust</i>	8	18.99	<i>entwickeln</i>	7	10.43	<i>Baum</i>	3	11.02	<i>Aussehen</i>	4	12.56
<i>genetisch</i>	14	18.44	<i>auswählen</i>	3	10.35	<i>Insel</i>	3	10.33	<i>Auswahl</i>	7	12.54
<i>tropisch</i>	9	18.19	<i>haben</i>	42	10.20	<i>Erde</i>	3	9.79	<i>Systematik</i>	3	11.85

Table 5: Word sketch for noun *Pflanze* ‘plant’.

Case	N	Conditions											
		incl. inflection						incl. inflection + affix-gender					
		1		3		5		2		4		6	
		R	P	R	P	R	P	R	P	R	P	R	P
Nominative	1,709	85	28	7	76	26	65	43	53	9	81	28	60
Accusative	618	64	24	6	37	18	41	51	30	6	35	14	45
Dative	149	62	9	21	34	41	35	55	13	25	59	40	74
Genitive	437	78	34			65	79	57	44			60	82

Table 6: Recall and precision in conditions 1-6.

5.6. Comparison with related work

To the best of our knowledge nobody has performed the identical task before. Closest to our work are probably German noun chunkers (Brants, 1999; Schmid and Schulte im Walde, 2000) and work on automatic extraction of German verb subcategorisation (Eckle, 1999; Wauschkuhn, 1999; Schulte im Walde, 2002).

Brants used Cascaded Markov Models to identify NP and PP chunks. He achieved up to 84.5% recall and 91.4% precision. However, he only evaluated the chunk structures but not the chunk labels, so it is difficult to compare his results with ours. Schmid and Schulte im Walde evaluated NP case labels in addition to the chunk structure, resulting in 83/84% recall and precision. Their results are considerably above ours, which can be attributed to their deeper approach. They used probabilistic context-free grammars.

Concerning verb subcategorisation acquisition, Eckle's methods were similar to ours, also being based on regular expressions over POS-tags and using morphosyntactic and structural constraints. Her objective was to obtain high-precision results, and she reported a recall of ca. 2%, with a precision of over 90%. Wauschkuhn's deeper, context-free grammar approach achieved a recall of 56.60% and a precision of 68.20%. The even deeper approach by Schulte im Walde achieved a recall of 69.74% and a precision 74.53%. Schulte im Walde (2008) provides an overview of approaches.

The differences in these results demonstrate how deeper approaches, usually requiring more input from linguists, result in higher precision and recall, whereas with shallow approaches, either precision or recall is sacrificed. Our approach is closest to Eckle's, in both method and results.

6. Discussion

6.1. Precision against recall

In the short term, our question is how to trade off precision against recall, to give the best output. Once the corpus and linguistic markup are given, the appropriate recall-precision tradeoff depends on users. As mentioned in the introduction, a user-oriented evaluation which may give some clues is currently under way. Different users may have different preferences, with respect to the application the word sketches are intended for. A lexicographer, for example, who relies on the availability of information that is as correct as possible (and who is potentially under time pressure to choose from the available word sketches), might prefer high precision over high recall. In contrast, a computational linguist who is interested in as much information as possible, even with low significance values, to avoid the sparse data problem, might prefer high recall over high precision. In sum, from a user's perspective, the German sketch grammar should be chosen according to the predominant application of the word sketches. Since we intend only to maintain one public version of the German sketch grammar, we shall be making the judgement according to the likely preferences of our "lead users", lexicographers and linguists studying lexis and grammar.

6.2. Corpus size

Working in our favour is the size of the corpus. In general the quality of a word sketch depends on the number of occurrences of a word in the corpus: a rule of thumb is that 600 instances give a good-quality sketch. So, for high-frequency words, even a small corpus will give a high-quality sketch, but for rarer words, the corpus needs to be very large for there to be high-quality word sketches for rarer words. Consider a word like *Waldbrand* 'forest fire'. The lemma has 1,976 instances in deWaC, and a healthy word sketch. In a BNC-sized corpus,⁵ one would expect around 120 instances: not enough for a good word sketch.

The rule-of-thumb of 600 is critically dependent on the recall of the sketch grammar. We may look at our word sketch for *Waldbrand*, find that it is too noisy, identify the looseness in the sketch grammar that allowed the noise to creep in, and add further constraints to improve the sketch grammar precision, at a cost to recall. Since *Waldbrand*, with 1,976, had ample data, it is likely that it will still have enough for a cleaner, higher-quality word sketch even after we have tightened the grammar. If, for example, we apply condition 6 with the effect of excluding around two thirds of the data from consideration when seeking subcategorisation patterns, we are still using a dataset of one third of 1,976 - over the 600 mark for *Waldbrand* and probably enough for a good word sketch.

We also note that word sketches are often much cleaner than recall and precision figures calculated over data instances (as above) might lead us to expect. The statistical sorting makes the system robust: even if there are many errors, they tend to be spread between different potential collocates, so, given large corpora, they rarely result in any single non-collocate getting a high enough salience score to appear in the word sketch.

6.3. The longer term

For the longer term, the message is simple. Richer linguistic analysis gives better performance.

A simple way to improve performance will be to integrate lexical lookup for finding noun gender, which will enable us to unambiguously establish noun case in more cases.

This will be part of a further strategy to prepare a POS-tagger which includes case, number and gender information in its output. The tagger might operate standalone, or as a post processor which takes as its input data already tagged by Tree Tagger using the STTS tagset, as in our experiments.

We are also looking into using parsers to pre-process the data. (Initial experiments have been undertaken for English.) While this paper has discussed sketch grammars, the Sketch Engine is also equipped to generate word sketches directly from parser output, without using a sketch grammar at all. Clearly, speed is a constraint: parsing 1.6 billion words is prohibitive for most parsers.

⁵The BNC, or British National Corpus, is a 100 million word corpus which is often used as a point of reference in corpus linguistics; <http://natcorp.ox.ac.uk>.

6.4. Lessons for other languages

We may also ask, how do our evaluation results transfer to other sketch grammars? Obviously, the German grammar is highly language-specific, so there cannot be any direct transfer of grammar rules (other than some trivial ones, such as adjective-noun combinations). However, we believe that the general methodology of integrating linguistic knowledge into the grammar rules to various degrees, in order to control for precision vs. recall of the grammar, is relevant to other languages as well. This general idea is not novel, of course, but this paper demonstrates how strong the effects of the grammar restrictions are, and thus how important it is to integrate and compare the various grammar parameters.

7. Conclusions

The paper set out to apply the Sketch Engine framework to German and to evaluate against manually annotated data. We selected German NPs as a test scenario, assuming that NP case is a crucial ingredient when extracting subcategorised verb complements and building useful word sketches. Furthermore, they are relatively easy to annotate, since in context they are easily disambiguated by the human reader.

Our sketch grammar defined NPs with respect to six linguistic conditions, taking inflectional and derivational information as well as sentence structure into account. As expected, if we restrict the rules, we get higher precision and lower recall. The recall is largest in the condition which puts the least restrictions on the sketch grammar definitions; the best precision is reached in the condition that incorporates inflectional and derivational information and in addition takes all clause types (and not only the clearly delimited type) into account.

The overall results are being evaluated by users in a separate study. Which German sketch grammar to choose (according to precision vs. recall values) should be decided according to the pre-dominant application of the word sketches.

The study has shown that the methods we have used are inferior to methods using richer linguistic inputs. This sets an agenda for us to improve German word sketches, by exploiting a lexicon to find noun gender, reviewing pos-tagging and in particular, the tagset we have been using, and, in the longer term, using richer parsing strategies.

8. References

- Marco Baroni and Adam Kilgarriff. 2006. Large Linguistically-processed Web Corpora for Multiple Languages. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy.
- Thorsten Brants. 1999. Cascaded Markov Models. In *Proceedings of 9th Conference of the European Chapter of the Association for Computational Linguistics*, Bergen, Norway.
- James Curran. 2004. *From Distributional to Semantic Similarity*. Ph.D. thesis, University of Edinburgh.
- Judith Eckle. 1999. *Linguistisches Wissen zur automatischen Lexikon-Akquisition aus deutschen Textcorpora*. Ph.D. thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.
- Stefan Evert. 2004. The Statistical Analysis of Morphosyntactic Distributions. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1539–1542, Lisbon, Portugal.
- Ulrich Heid and Marion Weller. 2008. Tools for Collocation Extraction: Preferences for Active vs. Passive. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, Marrakech, Morocco. To appear.
- Adam Kilgarriff, Pavel Rychlý, Pavel Smrz, and David Tugwell. 2004. The Sketch Engine. In *Proceedings of the 11th EURALEX International Congress*, pages 105–111, Lorient, France.
- Adam Kilgarriff, Chu-Ren Huang, Jan Pomikálek, Michael Rundell, Pavel Rychlý, Simon Smith, David Tugwell, and Elaine Uí Dhonnchadha. 2005. Word Sketches for Irish and Chinese. In *Proceedings of the Corpus Linguistics Conference*, Birmingham, UK.
- Simon Krek and Adam Kilgarriff. 2006. Slovene Word Sketches. In *Proceedings of the 5th Slovenian and 1st International Language Technologies Conference*, Ljubljana, Slovenia.
- Multext. 1995. Common Specifications and Notation for Lexicon Encoding and Preliminary Proposal for the Tagsets. Technical Report MULTTEXT Deliverable D1.6.1B, ILC, Pisa.
- Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen, 1999. *Guidelines für das Tagging deutscher Textcorpora mit STTS*. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, and Seminar für Sprachwissenschaft, Universität Tübingen.
- Helmut Schmid and Sabine Schulte im Walde. 2000. Robust German Noun Chunking with a Probabilistic Context-Free Grammar. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 726–732, Saarbrücken, Germany.
- Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging using Decision Trees. In *Proceedings of the 1st International Conference on New Methods in Language Processing*.
- Sabine Schulte im Walde. 2002. A Subcategorisation Lexicon for German Verbs induced from a Lexicalised PCFG. In *Proceedings of the 3rd Conference on Language Resources and Evaluation*, volume IV, pages 1351–1357, Las Palmas de Gran Canaria, Spain.
- Sabine Schulte im Walde. 2008. The Induction of Verb Frames and Verb Classes from Corpora. In Anke Lüdeling and Metja Kytö, editors, *Corpus Linguistics. An International Handbook*, Handbooks of Linguistics and Communication Science, chapter 44. Mouton de Gruyter, Berlin. To appear.
- Oliver Wauschkuhn. 1999. *Automatische Extraktion von Verbvalenzen aus deutschen Textkorpora*. Ph.D. thesis, Institut für Informatik, Universität Stuttgart.