Classifying NVG/FVG in an interactive parsing process

Alexander Geyken and Alexey Sokirko – Jan. 2006 – draft version

To appear in *Fellbaum, Christiane* (Ed.) : *Collocations and Idioms: Linguistic, lexicographic, and computational aspects.* London (Continuum Press)

Abstract

Studies concerned with the classification of German light and support verb constructions are commonly based on little or no empirical data. Now that very large corpora are available, however, exploiting them is difficult since nearly all verbs that participate in light and support verb constructions are too frequent to allow a manual evaluation of all the corpus examples of constructions headed by these verbs. We propose a semi-automatic method where the human lexicographer is supported by a shallow parser. Our goal is to accelerate the syntactically based classification of the constructions.

Introduction

Eisenberg (1999) calls noun-verb collocations like in Erscheinung treten ('make [one's] appearance') and eine Entscheidung treffen ('make a decision') Nominalisierungsverbgefüge or verb-nominalization constructions (NVG). A related class of constructions called Funktionsverbgefüge or function verb constructions (FVG) is usually considered a subclass of NVG (von Polenz 1987, Eisenberg 1999). For FVG, the meaning of the entire construction has specific grammatical properties that are absent in the corresponding simplex verb, e.g., a change of aspect or passivization (see Storrer, this volume). For the most part, FVG have the structure verb + preposition + predicative noun; hence, the example given above in Erscheinung treten is an FVG, whereas eine Entscheidung treffen is an NVG but not an FVG. To distinguish FVG and NVG linguists have considered the syntactic restrictions (choice of determiner, attribution, pronominalization, etc.), the degree of grammaticalization or lexicalization, the relation between the NVG/FVG and the base verb or the classification of NVG/FVG according to formal and informal communication style.

Up to now, these questions have largely been investigated in the absence of empirical evidence. Grammatical research has proceeded on the basis of constructed examples focusing on typical cases intended to illustrate the most important properties of a construction. Even now, linguists generally use corpora only to extract data that confirm a given point or theory. An example is the extensive work of van Pottelsberge (2001) who uses corpus data but limits his selection largely to those that support his refutation of the criteria for distinguishing NVG and FVG proposed in the literature; van Pottelsberge does not offer an unbiased qualitative and quantitative analysis of the phenomena. However, some recent work has used corpora to address the questions more broadly. Thus Storrer (2006) demonstrates in a corpus-based study a range of differences between NVG and the corresponding base verbs (e.g., Unterricht erteilen 'give instructions', and unterrichten 'instruct') that can only be discovered and described when naturally occurring data and larger contexts are available. Eisenberg (2003) shows on the basis of a corpus-based analysis with legal texts written between 1706 and 1995 (Seifert 2004) that the proportion of FVG declines continuously after the beginning of the 19th century, while the number of NVGs (that are not FVGs) increases. Furthermore, the

distribution of the FVG construction becomes more restricted and limited to very few lexical units, an observation that argues in favor of a lexicalization tendency of FVG. On the other hand, the growing number of nouns in the NVGs can be interpreted as one possible cause for a general nominalization tendency in German.

In both these studies, corpus data were manually evaluated. In particular, it was necessary to count the number of cases where a co-occurrence of a noun and a verb constituted an FVG or NVG in order to determine typical syntactic constraints and to better understand the semantic nature of the nouns in NVG/FVG.

For small or medium-size corpora, manual inspection of the data is a considerable but feasible task. But such analyses cannot be extended to the large corpora containing billions of words that are now available. For example, the verb leisten ('render', 'perform') occurs 98,208 times in the one-billion-word DWDS corpus (Geyken, this vol.); halten ('hold') occurs 488,185 times, and a search for kommen ('come') yields 1,123,192 hits (a more detailed list can be found in Appendix 1). A manual analysis of the KWIC lines generated by verb-based queries is thus impossible. A solution commonly adopted by lexicographers is sampling, i.e., the random selection of every n^{th} corpus example that cuts down the data to a manageable size. An obvious disadvantage of this method is that reducing the data also reduces their variety and thus may impair the quality of the analysis. A much better approach is filtering, where examples are automatically pre-classified in ways specified by the linguist. Such prior classification should be able to classify V-N constructions syntactically.

Prior classification should be robust in the sense that it must return an analysis for each sentence. Precision is a priority, complete recall less so. This corresponds to the expectation of the lexicographer that the process correctly classifies the examples even if it cannot classify a number of examples at all. Moreover, prior classification should be an interactive process, i.e., the lexicographer should be able to change the queries or complete them in a way that increases recall.

A simple experiment shows that POS tagged and lemmatized corpora are insufficient for our purposes. By way of example, we extracted from the DWDS 500 sentences where the strings Dienst ('service') and leisten ('render', 'perform') co-occur.

Approximately 80% of the hits are examples of the NVG Dienst leisten ('render/perform a service'). In 20% of the cases, Dienst is part of a phrase with a different case marking, e.g. im Dienste, or it is the subject of die Dienste leisten gute Arbeit ('the services perform solid work'), or it occurs within another clause in the case of complex sentences. Examples will be given below.

It follows that a satisfactory extraction must recognize at least phrases (so as to distinguish prepositional phrases from simple nominal phrases) and clauses (so as to distinguish dependent N-V constructions from dependencies that happen to co-occur in the same sentence).

Several parsers developed specifically for German meet these criteria: KaroPars (Ule & Müller 2004, Müller 2004), LoPar (Schmid, 2000), the shallow parser of (Neumann & Piskorski, 2002), and the parser Synan developed in house (Sokirko 2004). As the focus of this chapter is not a comparison of different parsers but rather the challenge of classifying data in very large corpora, we used the in-house parser Synan. An example is given in Figure 1.

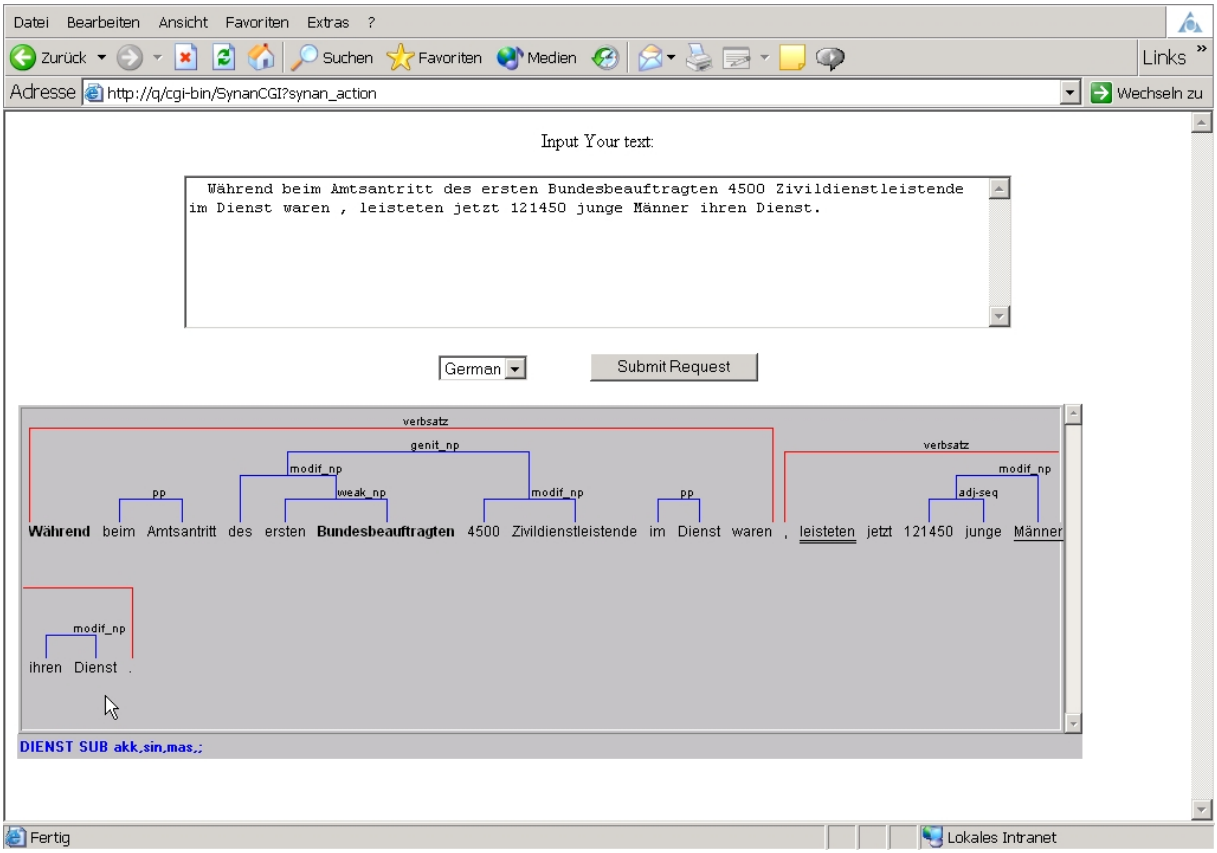

Fig. 1

The sequences marked in blue are phrases that constitute lexical categories and operate on the phrasal level. The main verb of the matrix clause is underlined twice, the subject of the matrix clause once. The sequences marked in red are clausal categories that define possible clause constituents, based on punctuation.

Clauses are built via special clausal rules, which are applied iteratively to the entire sentence. The process stops when the application of the rules yields no new phrases or clauses. The result of the parse process is a set of parse trees ordered according to relevance (coverage) that assign to each word its maximally disambiguated morphological interpretation and—when possible—assign it to a phrase or a clause.

The interactive parser

In this process, the shallow parser receives as input an example corpus and a set of verified verb-noun constructions, for example all those that are listed in a dictionary for a given verb. The parser breaks down the sentences into clauses and phrases. An evaluation script then extracts all verb-noun constructions, marks them in the examples and classifies them according to syntactic patterns such as passivization, infinitive construction, etc. The remaining examples are likewise pre-classified syntactically and passed on to the lexicographer for validation. The lexicographer identifies additional verb-noun constructions, and these are re-analyzed by the parser and classified. In this way the list of examples identified by the lexicographer as non-NVG is steadily reduced until the parser cannot

identify any new examples or until no more relevant hits are found in the example corpus. The result of this iterative process is an example corpus classified according to syntactic patterns.

Two small-scale experiments were carried out to check the level of accuracy with which the parser pre-classifies sentences. Below we sketch the method for prior classification, describe the experiments, and evaluate our methods with respect to recall and precision and fine-grainedness of the analysis. Finally, we discuss the extensibility to different FVG and NVG.

Method

The following procedure is used to classify the corpus example sentences.

Step 1: Extraction of candidate sentences with FVG and NVG from the DWDS corpus by means of a DDC query (Geyken, this vol.). A superset of FVG is extracted with a Boolean query.

Step 2: The lexicographers issue a list of the FVG and NVG to be extracted, whose nouns are listed in the dictionary under the verb entry as verb-noun collocations. This list is encoded in a machine-readable format (cf. Appendix 2).

Step 3: All corpus example sentences are parsed with Synan.

Step 4: Classification of the analysis via an evaluation script that distinguishes the following cases:

a. If verb and noun are in the same clause and the noun is properly case-marked then classify the sentence as NVG/FVG.

b. If verb and noun are not in the same clause or if the noun is not properly case-marked then classify the sentence as “No result”.

Moreover, the evaluation script marks the NVG or FVG in the example and annotates it with syntactic tags like passive, infinitival construction, relative clause, etc.

Experiment 1

This experiment evaluates our method. Steps 1–4 (described above) were applied to the NVG Dienst leisten (‘render/perform [a] service’). First, we extracted from the one-billion-word DWDS corpus all cases of co-occurrences of the strings Dienst and leisten. We randomly selected a set of 500 examples from the 4,600 hits that were returned. This example set was manually inspected for cases where the noun Dienst and the verb leisten occurred as NVG. For cases where the co-occurrence of these two strings did not constitute an NVG, we noted whether leisten formed an NVG with another noun. The resulting annotated corpus constitutes the training corpus for the parser, i.e. the outcome of the parse was evaluated against the annotations in the training corpus; some of the rules of the parser had to be adjusted and improved in several cycles. Finally, we obtained the following results.

Results

In 418 of the 500 returns, Dienst leisten constituted an NVG. In the remaining 82 sentences, Dienst und leisten co-occurred in the same sentence, but either in a different construction (examples 1 und 2), with Dienst in subject position (3), or in separate clauses (4):

(1) In einem zunehmenden, auf Wettbewerb ausgerichteten Markt könnten es sich Telekommunikationsunternehmen, Betreiber von Sendern, Kabel-TV-Gesellschaften oder Telefondiensteanbieter nicht **leisten**, den Markt für interaktive **Dienste** etwaigen Kunden vorzuenthalten.

'In a growing and competitive market, companies offering telecommunication, broadcast, cable TV, as well as telephone service providers cannot **afford** to withhold interactive **services** from potential customers.'

Interaktive Multimediadienste für Kunden einfach gestalten, in: Frankfurter Allgemeine Zeitung, 10.12.1996, S. 10

Note: leisten here occurs in an unrelated sense, 'afford'.

(2) Wie aber will Gasser einen "systematischen Vergleich beider Theorienkomplexe" **leisten**, wenn auf der einen Seite ein im Bau befindliches wissenschaftliches Gebäude steht, während auf der anderen Seite Aphorismen mehr und mehr die rohe Gestalt von Gesetzestafeln annehmen , die Nietzsche im **Dienst** einer höheren Sache als der Wissenschaft meißelt?

'How does Gasser want to **perform** a "systematic comparison of both theory complexes," when a science building is currently under construction on the one hand, while on the other hand aphorisms increasingly take on the form of law tablets that Nietzsche carves in the **service** of something higher than science?'

Der Wille zum Buch, in: Frankfurter Allgemeine Zeitung 21.04.1998, S. 45

(3) Eine unannehmbar große Zahl von Kenianern lebe unterhalb der Armutsgrenze und habe keinen Zugang zu **Diensten**, die die Grundversorgung **leisteten**.

An unacceptably large number of Kenyans live below the poverty line and have no access to **services** that **provide** basic care.

[Am 6. September ...] [03.10.99], in: Archiv der Gegenwart 69 (1999), S. 43822)

(4) Diese **Dienste** dürften so billig sein, daß sich in Amerika jede Schule und jedes Krankenhaus eine Verbindung werde **leisten** können.

‘These **services** should be so inexpensive that every school and every hospital in America should be able to **afford** a connection.’

„Das Ausmaß der Informationsrevolution wird noch immer unterschätzt“, in: Frankfurter Allgemeine 07.12.1998, S. 13

Note that leisten here means ‘afford’ rather than ‘perform.’

In 393 of 418 cases the evaluation script identifies Dienst and leisten as occurring in the same clause and with the noun marked in the accusative case. This is correct in 392 cases, and only one sentence is erroneously analyzed as an NVG. This corresponds to a recall of 93.7% and a precision of 99.7%. Furthermore, the syntactic analysis yields the following classification.

(i) Active, finite verb form	293
(ii) Noun not in NVG use	42
(iii) Infinitive clause	42
(iv) Passive	22
(v) Modal infinitive	4

The following examples illustrate these cases.

(i) Active, finite verb form:

(5) Man könne im Namen des deutschen Katholizismus erklären, daß der österreichische Hirtenbrief in historischer Stunde der Schicksalsgemeinschaft des gesamtdeutschen Volkes einen schlechten **Dienst geleistet habe**.

‘One could declare in the name of German Catholicism that the Austrian pastoral letter in the historic hour **did a disservice** for the entire German people who shared the same fate.’

o.A., Nationalsozialismus. Kirchen, Katholizismus, Ständischer Aufbau. Verschiedenes, Jugendverbände, Beziehungen zu ÖSTERREICH [14.01.34], in: Archiv der Gegenwart 4 (1934), S. 1226

(ii) Conversion to adjective noun phrase:

(6) Das Kabinett wurde mit Dank für die **geleisteten Dienste** mit der vorläufigen Weiterführung der Geschäfte betraut.

‘The cabinet was thanked for the **services rendered** and entrusted with the continued leadership of the affairs.’

o.A., Reichsregierung, Reichspräsident [28.01.33], in: Archiv der Gegenwart 3 (1933), S. 672

(iii) Infinitive clause:

(7) Die rumänische Regierung erließ am 17. September folgendes Gesetz: In Kriegszeiten sind die Frauen verpflichtet, dem Vaterlande **Dienste zu leisten**.

‘The Romanian government issued the following law on September 17: In war times, women are obliged to **render services** to the fatherland.’

o.A., Wehrmacht [04.10.38], in: Archiv der Gegenwart 8 (1938), S. 3750

(iv) Passive:

(8) Ende Dezember 1932 hat Bulgarien mit seinen ausländischen Gläubigern ein Übereinkommen geschlossen, laut welchem der **Dienst** der ausländischen Anleihen bis zum 31. März 1933 zu 40% (bisher 50%, s. 479 O) in Devisen **geleistet wird**.

‘In late 1932 Bulgaria reached an agreement with its creditors, according to which the **service** of the foreign loans will be **performed** in foreign currencies until March 31, 1933 at 40% (previously 50%, cf. 479 O).’

o.A., Verschuldung [03.01.33], in: Archiv der Gegenwart 3 (1933), S. 635

(v) Modal passive (and relative clause):

(9) Zu persönlichen **Diensten**, die grundsätzlich ohne Vergütung **zu leisten sind**, werden Personen nicht herangezogen, die wegen ihres Lebensalters oder ihres Gesundheitszustandes nicht geeignet sind.

‘People who are disqualified for age or health reasons will not be asked to perform personal **services** that in principle **are to be rendered** without compensation.’

o.A., Vorschläge U Thants für intensivere Aktionen der Vereinten Nationen während des „Vereinte Nationen-Entwicklungsjahrzehnts“ [01.07.62], in: Archiv der Gegenwart 32 (1962), S. 9955

Experiment 2

In a second experiment we selected another 100 examples from the example corpus, making sure that no data used in Experiment 1 were re-used. The goal of this control experiment was to check whether the results of Experiment 1 were corrupted as a result of the fine-tuning of the parse rules.

Results

The manual evaluation of the 100 sentences showed that 78 examples were NVG. The syntactic analysis identified 76 of the 78 examples as the NVG Dienst leisten; this classification was correct in 74 cases and incorrect in only 2 cases. This corresponds to a recall of 94.8% and a precision of 97.3%.

Similar to the outcome of Experiment 1, the syntactic analysis yielded the following classification:

(i) Active, finite verb form	65
(ii) NP not in NVG use	6
(iii) Infinitival sentence	1
(iv) Passive	4
(v) Modal Infinitive	0

Of the remaining 22 examples, 16 contain additional NVG. Assuming that the lexicographer has included all nouns listed in the entry for leisten in the WDG dictionary (www.dwds.de/wdg) in his initial set of NVG candidates involving the verb leisten, the parsing process can identify the following tokens among the 16:

- Recognized in the dictionary and by the parser: Folge(‘suit’), Hilfe (‘help’), Widerstand (‘resistance’).
- Not in the dictionary and thus not recognized by the parser in the first iteration: Beitrag (‘contribution’) (2), Beistand(‘assistance’), Daseinsvorsorge (‘government provision of services for basic survival’), Dienst ableisten (‘render services that are owed’), Einiges (‘quite a bit’) (2), Infrastrukturaufwand (‘effort needed for preserving the infrastructure’), Pflege (‘care’).
- In the second iteration, these would be added to the NVG list and recognized as a possible complement with accusative case marking in the same clause. Fig. 2 shows an example.

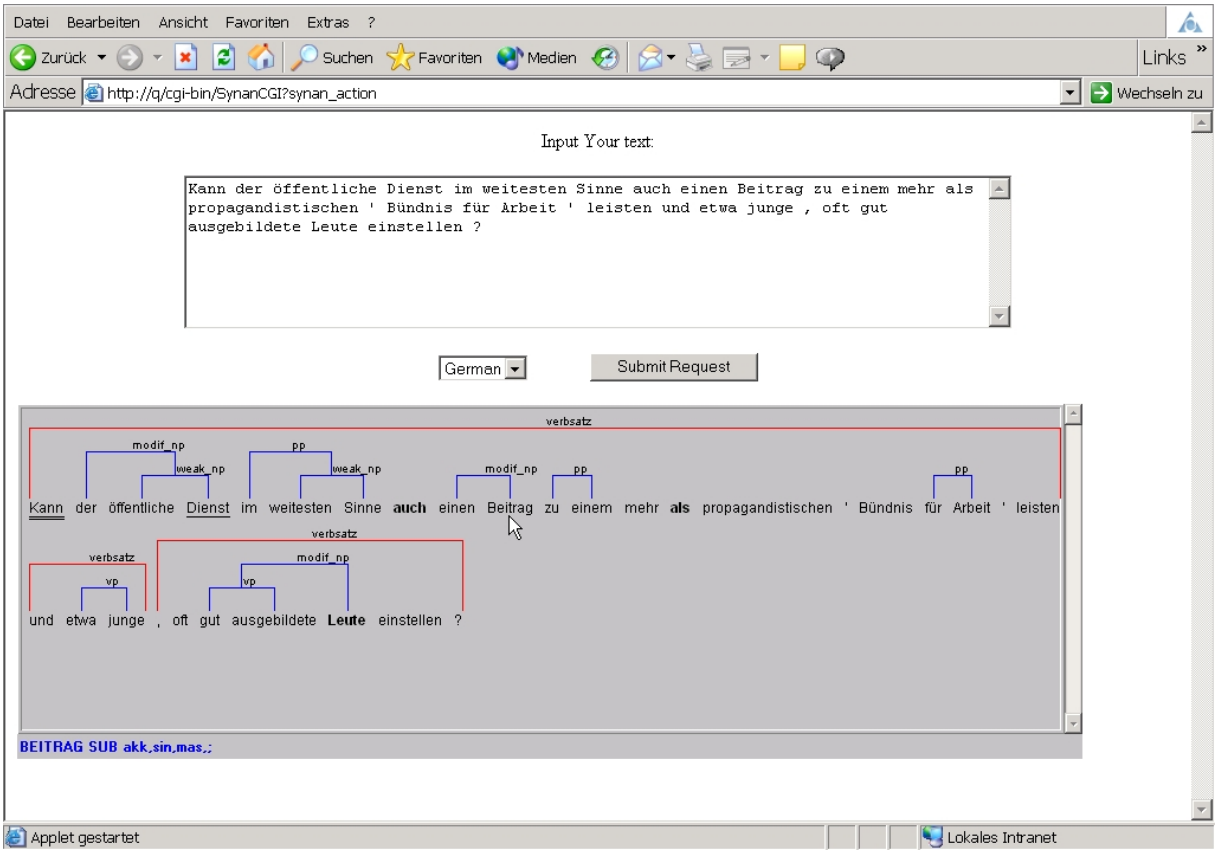

Fig. 2

The following cases are syntactically too complex for the current version of the parser.

(10) Mögen auch von dieser feierlichen Stunde der Eröffnung des ersten regelmäßigen Fernsehprogramm**dienstes** an noch Tausende Hindernisse zu überwinden und Jahre voll härtester Arbeit **zu leisten sein**, wir werden dieses Ziel erreichen, im Dienste an unserer stolzen deutschen Volksgemeinschaft.

'Even if, after this festive hour when we open the first regular television broadcasting **service**, there are still thousands of obstacles to overcome and years of hard work to be **performed**, we will reach our goal in the service of our proud German people.'

o.A., Luftschutz [27.06.35], in: Archiv der Gegenwart 5 (1935), S. 2111

(11) Im Öffentlichen **Dienst** erhält vergleichsweise ein Festangestellter mindestens 60 Mark brutto die Stunde und muß nur die Hälfte der Sozialabgaben **leisten**.

'An employee in the public **service** receives at least 60 Marks an hour and must **pay** only half of that for social security deductions.'

„Die freien Mitarbeiter sind dem Arbeitgeber völlig ausgeliefert“, in: Süddeutsche Zeitung 17.01.1996, S. 39

(12) Die neue Finanzierungssituation führe nämlich dazu, daß professionelle **Dienste** künftig nur noch 'ein Minimum' an Versorgung **leisten** könnten, erklärte Marianne Winter vom Beratungsdienst der Caritas.

‘The new financing scheme would lead to a situation where professional **services** could only **provide** a minimum of care in the future, declared Marianne Winter of the counseling service of Caritas.

Caritas-Zentrum in Ramersdorf: Aus für den Mobilen Sozialen Hilfsdienst, in: Süddeutsche Zeitung 31.08.1995, S. 1

(13) Als Spieler und als Trainer **hat** er viel für Deutschland **geleistet** und sich immer in den **Dienst** des Ganzen gestellt.

‘As a player and as a coach he **accomplished** much for Germany and he always stood in the **service** of the larger goal.’

Leute von heute, in: Süddeutsche Zeitung 06.07.1996, S. 12

Note: leisten has a sense ‘accomplish’ unrelated to the NVG use of leisten.

Discussion

For the task of identifying the NVG Dienst leisten, a precision of 99.7% (Experiment 1) and 99.7% (Experiment 2) is significant better than the precision one would achieve with a purely Boolean query, which yields only 81% precision. Precision here is crucial, as the lexicographer does not want to inspect pre-classified examples, which would wipe out all savings of time and effort.

Similarly, a recall of 93.7% (Experiment 1) and 94.8% (Experiment 2) is very high. The cases that are not recognized are sentences containing either a different NVG or a syntactic construction that is too complex for a shallow parser. In the first case the parser can correctly assign 75% of the cases provided that the nouns were marked by the lexicographer as those that can occur in NVG. In the following example the parser would thus correctly recognize the NVG Beitrag leisten (‘make a contribution’) as Beitrag is the only complement marked with the accusative case:

(14) Der Staat **leistet** für den **Dienst** dieser Obligationen während der gesamten Laufzeit einen Beitrag von jährlich höchstens 2,25 %.

‘The state **makes** an annual contribution of no more than 2.25% for the **service** of these obligations during their entire maturity.’

Landwirtschaft, Schuldenregelung, Finanzen [13.10.35], in: Archiv der Gegenwart 5 (1935), S. 2257

The limits of shallow syntactic analysis are reached in the case of complex phrasal structures such as aggregate functions (e.g. (12)) or coordination and anaphors whose resolution requires semantic criteria or scope (e.g. (15)).

(15) Auch die ausführliche Dokumentation der europäischen Rechtsakte und Dokumente am Ende des zweiten Bandes sowie das umfangreiche Rechtsprechungsverzeichnis **leisten** gute **Dienste**.

‘The extensive documentation of the European legal files and the documents at the end of the second volume as well as the comprehensive legal register **provide** useful **services**.’

Umweltrecht zum Schmökern, in: Frankfurter Allgemeine 17.01.2000, S. 14

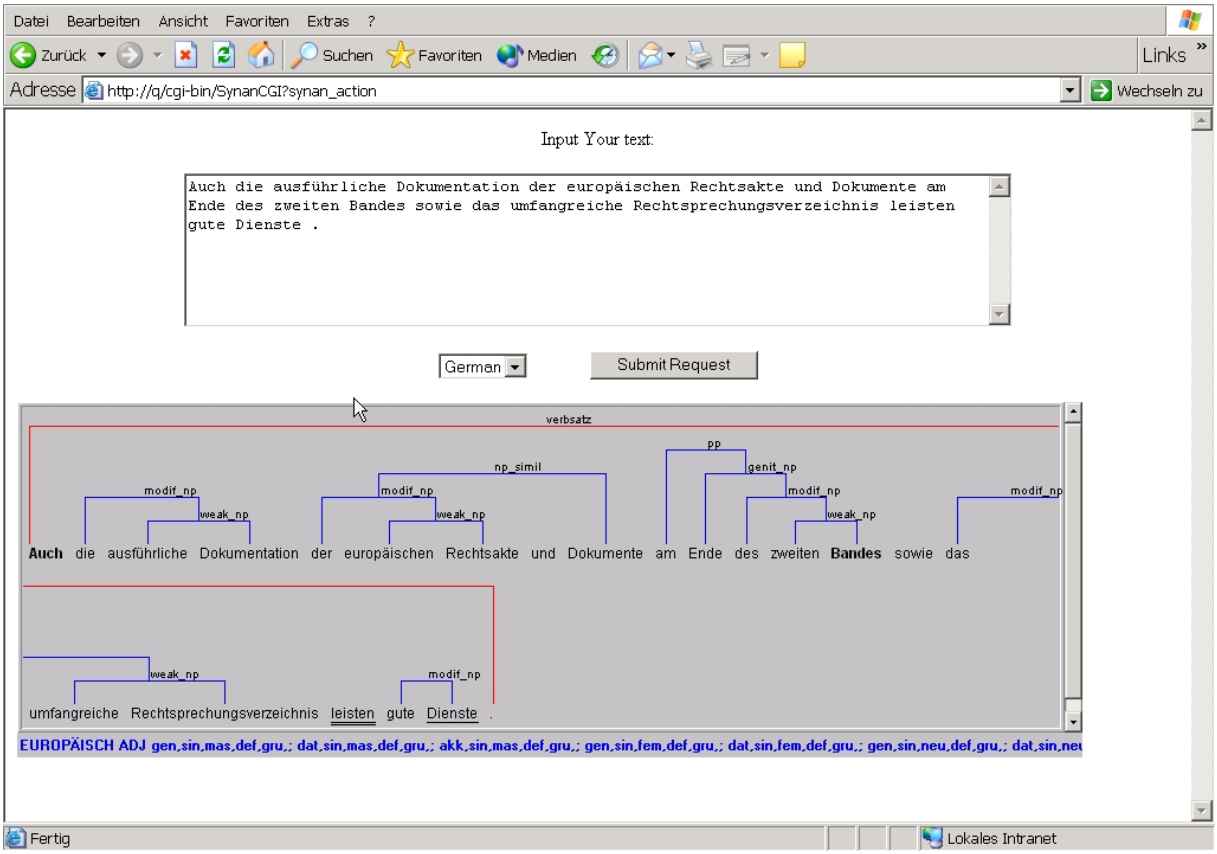

Fig. 3

In this example, ‘extensive documentation of the European legal files and the documents at the end of the second volume as well as the comprehensive legal register’ is not analyzed as a

coordinated structure. Thus, Dienste is the only plural noun and is erroneously identified as the subject because of its agreement with the verb.

A flat syntax appears to be optimal analysis for successful classification. Pure bottom-up chunking does not suffice for complex adjective phrases (cf. (16)). On the other hand, a deep parser would not be robust enough and would not render a complete analysis for complex sentences (cf. also (16)). The shallow parser cannot resolve all dependencies in the sentence but can determine that leisten and Dienst are clause mates and that Dienst is marked with the accusative case.

(16) Auf Grund einer weiteren Verordnung (Goldschuldenerleichterungsverordnung) ist der **Dienst** der von österreichischen Instituten ausgegebenen, auf Schilling mit Goldklausel lautenden Pfandbriefe und der fundierten Bankschuldverschreibungen, sofern die bezüglichen Zahlungen nach dem 30. April 1933 fällig werden, unter Zugrundelegung des von der Wiener Börsekammer errechneten Goldpreises **zu leisten**.

'As a result of an additional regulation (regulation of the easement of gold debts) the **service has to be delivered** for covered bonds, nominated on Austrian Schillings with a gold clause issued by the Austrian institutes, and for grounded Bankschuldverschreibungen, if those payments are due after April 30th, 1933 on the basis of the gold prize fixed by the Vienna stock exchange.'

Währung, Zinsensenkung [24.03.33], in: Archiv der Gegenwart 3 (1933), S. 760

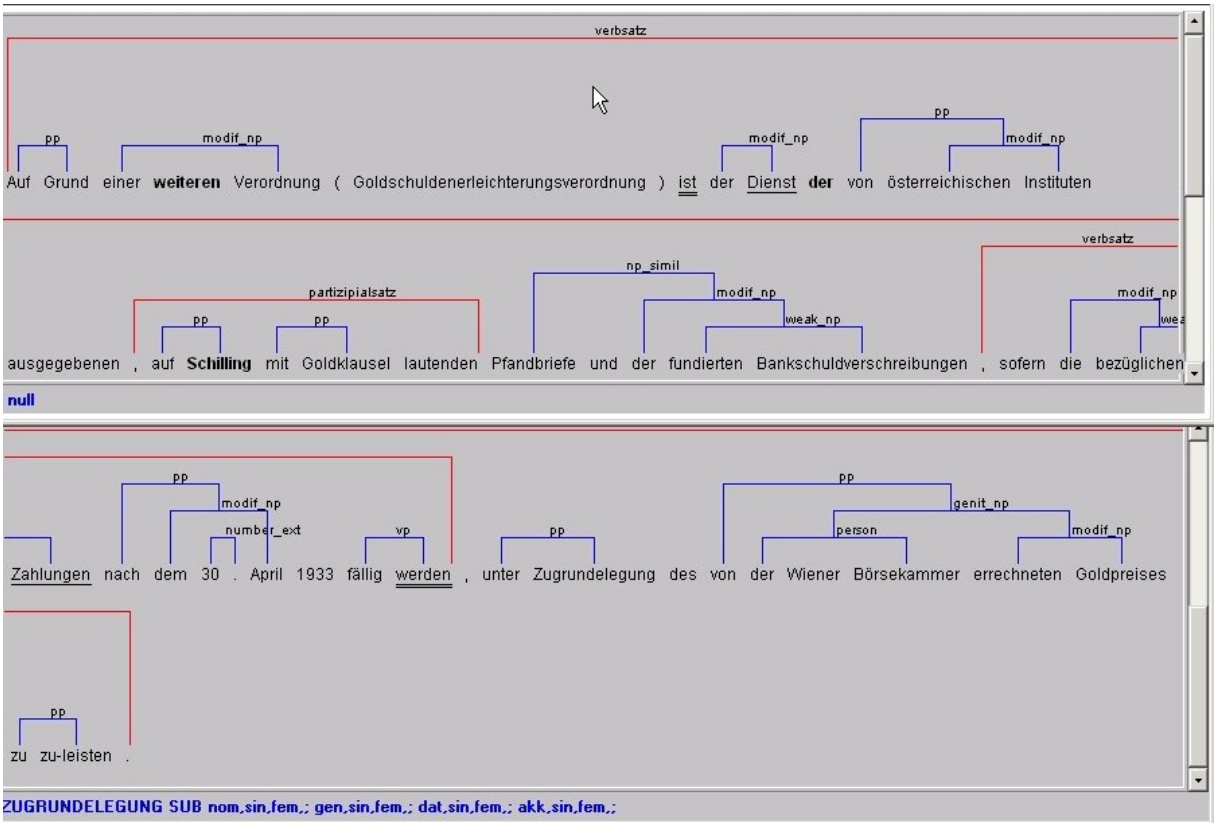

Fig. 4

In concluding we can say that this study yields encouraging results for the general case of automatic prior classification of all NVG/FVG with a given verb. The parser must first extract a noun candidate. Nouns can occupy different complement positions, and identifying a prepositional complement is much harder than an accusative object, as the evaluation of KaRoPars shows (cf. Müller 2004). Hence it is to be expected that recall and precision of the prior classification will be lower than in the case of Dienst leisten.

Bibliography

Eisenberg, P. (2003): Funktionsverbgefüge in Gesetzestexten: Mittel der Distanzkommunikation. Written version of a lecture delivered at the symposium 'Kommunikation im Recht' organized by the Berlin Brandenburg Academy of Sciences, October 4, 2003, Blankensee.

Eisenberg, P. (1999): Grundriß der deutschen Grammatik. Band 2: Der Satz. Stuttgart/Weimar: Metzler.

Klein, Wolfgang, Alexander Geyken (2000): 'Projekt "Digitales Wörterbuch der deutschen Sprache des 20. Jh."'. In: Jahrbuch der BBAW 1999, Berlin: Akademie Verlag, S. 277–289.

Müller, F. H. (2004): 'Annotating grammatical functions in German using finite-state cascades'. In: Proceedings of the Twentieth International Conference on Computational Linguistics (COLING 2004). Genf, Schweiz, August 2004.

Neumann, G. and Piskorski J. (2002): 'A shallow text processing core engine'. In: Journal of Computational Intelligence, Volume 18, Number 3, 2002, 451–476.

Polenz, Peter von (1987): ‚Funktionsverben, Funktionsverbgefüge und Verwandtes. Vorschläge zur satzsemantischen Lexikographie'. In: Zeitschrift für germanistische Linguistik. 15. 169–189.

Schmid, H. (2000): LoPar: Design and implementation (Unpublished ms., Institute for Computational Linguistics, University of Stuttgart).

Seifert, J. (2004): Funktionsverbgefüge in der deutschen Gesetzessprache (18.-20. Jahrhundert). Hildesheim: Olms.

Sokirko, A. (2004): A Shallow Syntax System for DWDS. Technical Report. www.dwds.de/dokumente/GerSynan.pdf

Storrer, A. (2005): ‚Funktionen von Nominalisierungsverbgefügen im Text: eine corpusbasierte Fallstudie'. In: Proust, C., Winkler, E, (Hg.). Festschrift für Gisela Harras. Tübingen: Narr 2006.

Ule, Tylman und Müller, Frank Henrik (2004): ‚KaRoPars: Ein System zur linguistischen Annotation großer Text-Korpora des Deutschen'. In: Automatische Textanalyse. Systeme und

Methoden zur Annotation und Analyse natürlichsprachlicher Texte. Hg. Alexander Mehler und Henning Lobin. Opladen: VS Verlag für Sozialwissenschaften.

Van Pottelsberge, J. (2001): Verbonominale Konstruktionen, Funktionsverbgefüge. Vom Sinn und Unsinn eines Untersuchungsgegenstandes. Heidelberg: Winter.

Appendix 1: Frequency of some NVG/FVG verbs in the DWDS Corpus.

Lemma	**Frequency in DWDS Corpus**
sein (‘be’)	14,460,677
haben (‘have’)	8,910,664
geben (‘give’)	1,412,542
kommen (‘come’)	1,123,192
gehen (‘go’)	1,000,011
machen (‘make’)	929340
stehen (‘stand’)	903084
bleiben (‘remain’)	653906
liegen (‘lie’)	567280
finden (‘find’)	543067
nehmen (‘take’)	455969
stellen (‘put’)	449103
bringen (‘bring’)	376398
treten (‘step’)	202149
treffen (‘meet’)	184219
befinden (‘feel’)	132791
leisten (‘provide, make’)	98208
erteilen (‘grant’)	43227
üben (‘exercise’)	34856

Table 1: Verb frequencies in the DWDS Corpus

Appendix 2: Machine-readable encoding of dictionary entries

This appendix gives an example of the machine-readable format of the two senses of Dienst leisten in the WDG dictionary.

Sense 1: jmd. leistet jdm. einen Dienst
<PATTERN id="leisten_Dienst01"
<NP syn="Subj",case="nom",sem="HUM" type="m"/>
<VP lemma="$leisten"/>
<NP syn="DirObj",case="acc",wf="$Dienst" type="m"/> .
<NP syn="IndirObj",case="dat",sem="HUM" type="o"/> .
</PATTERN>

Example: Nur auf diese Art können wir Frankreich einen wirklichen **Dienst leisten**.
'Only in this way we can **provide** a true **service** for France.'

Sense 2: etw. leistet jdm. einen Dienst
<PATTERN id="leisten_Dienst02"
<NP syn="Subj",case="nom",sem="-HUM" type="m"/>
<VP lemma="$leisten"/>
<NP syn="DirObj",case="acc",wf="$Dienst" type="m"/> .
<NP syn="IndirObj",case="dat",sem="HUM" type="o"/> .
</PATTERN>

Example: Ein Deichselträgerrad **leistet** fast gleiche **Dienste**.
'An axle carrier wheel **provides** almost the same **service**.'

Notation:
The elements NP, VP stand for Noun Phrase, Verb Phrase.
id = unique key for a corpus pattern
syn = (Subj | DirObj | IndirObj)
type = (m | o) /* m=mandatory, o=optional
case = (nom | acc | dat | gen)
wf = #pcdata /* (wf=wordform
lemma = #pcdata /* $ denotes the lemma operator
sem = {THES} , where THES corresponds to a shallow semantic hierarchy such as: HUM=human, ABSTR=abstr, SCHADEN=harmfulness
Lexical categories are denoted in STTS tagset: i.e. APPR, NN, ...