Design and application of a Gold Standard for morphological analysis: SMOR as an example of morphological evaluation

Gertrud Faaß, Ulrich Heid, Helmut Schmid

Universität Stuttgart
Institut für maschinelle Sprachverarbeitung
Azenbergstr. 12
D 70174 Stuttgart
faasz,heid,schmid@ims.uni-stuttgart.de

Abstract

This paper describes general requirements for evaluating and documenting NLP tools with a focus on morphological analysers and the design of a Gold Standard. It is argued that any evaluation must be measurable and documentation thereof must be made accessible for any user of the tool. The documentation must be of a kind that it enables the user to compare different tools offering the same service, hence the descriptions must contain measurable values. A Gold Standard presents a vital part of any measurable evaluation process, therefore, the corpus-based design of a Gold Standard, its creation and problems that occur are reported upon here. Our project concentrates on SMOR, a morphological analyser for German that is to be offered as a web-service. We not only utilize this analyser for designing the Gold Standard, but also evaluate the tool itself at the same time. Note that the project is ongoing, therefore, we cannot present final results.

1. Introduction

1.1. Perspectives on evaluation

There are usually two perspectives to be considered when NLP tools are evaluated: the developer's and the users' view. Developers validate their tool by comparing the input/output pairs to what they expect, but they also check e.g. for the processing speed or other system parameters. Such validation of specific targets by the developer is dependent on the system's knowledge base (e.g. lexicon contents and processing rules), in other words, developers validate and report on the performance of their system on the basis of what they expect it to be capable of doing.

From the users' perspective, system performance has to satisfy their requirements. We refer to Underwood (1998) who states – for NLP lexicons – that users' requirements may significantly differ when being compared to what a system has to offer; this ranges from needing far less information than what the system has to offer to needing to extend or modify even the best output. Additionally, in the light of an increasing number of web services offering linguistic analysis (including morphological analysis), the user should have the possibility to compare between different tools on offer.

When taking both requirements into account, the EAGLES recommendations (EAGLES, 1996) are of use: these state that any NLP system should be evaluated – and measured – according to a given list of features that translate the requirements of a user group, thus making systems comparable.

Hence, there are actually three evaluation perspectives, i.e. developer's aims, user's requirements and the possibility of comparing between systems. Based on these perspectives, we can formulate several basic questions that should be answered in the course of an evaluation:

(1) Does the tool do what it is supposed to do (developer's validation)?

(2) Which features does the system offer (the qualitative aspect)?

(3) How big are the system's knowledge sources, i.e. how good is its coverage (the quantitative aspect)?

(4) Can (2) and (3) be represented as a list of attribute-value pairs in order to enable users to compare different systems?

When these questions are applied to a system for inflectional morphological analysis, evaluation aspects will concern inter alia the correct assignment of base lemmas and parts of speech (henceforth POS), the correctness and completeness of inflectional paradigms and of compounding and derivation, and the system's coverage when analysing random word forms.

Referring to Underwood (1998), we do not expect the user to have to reduce (or post-process and/or add to) a tool's output in terms of the coverage or of the granularity of the analyses it provides. There hence will be different version of the Gold Standard necessary to represent the needs of different user groups appropriately, i.e. the several versions will differ with respect to the granularity of analysis: all will contain the same, if possible **near-representative**, list of word forms and their expected analyses at a certain level of granularity. A noun like *Priesterweihe* ("Priest ordination"), for example, might be analysed in the different versions of the Gold Standard as a singular and feminine noun or, additionally, as a compound of the two nouns *Priester* and *Weihe*, or even as a compound of the noun *Priester* and the derived noun *Weihe* (which has the verb *weihen* ("[to] ordinate") as a basis), cf. (3) in paragraph 4.2.2.

A system's ability to deliver analyses identical to the ones proposed by one or several Gold Standard version(s) is measurable; evaluation results gained here will serve as a basis for a commensurable documentation: a comparison of several systems is hence made possible.

This paper describes the design and application of a Gold Standard for morphological analysis in general and some preliminary results. Our work is part of the "D-Spin" project[1] (cf. e.g. (Bankhardt, 2009)), where it is planned to make the morphological analyser SMOR (Schmid et al., 2004) available as a web service; it is also planned to provide detailed documentation on its quality and coverage. Additionally, a systematic evaluation of the morphological rules currently used by the system are carried out in order to further enhance its quality.

1.2. SMOR as part of a web service

SMOR (Schmid et al., 2004) is a German finite-state morphology covering inflection, derivation, and compounding which was implemented with the SFST tools (Schmid, 2005). SMOR uses the IMSLex lexicon (Fitschen, 2004) and contains a comprehensive list of 47,671 German base stems, 528 compounding and 1,691 derivation stems[2] as well as 323 prefixes and 208 suffixes. With these morpheme entries, SMOR is able to analyse a much larger number of word forms. The morphological units are labeled with features which are described in (Schmid et al., 2001). Features of stem entries specify the part of speech, the stem type (base, derivation, or compounding stem), the origin (native, foreign, and several subtypes of neoclassical stems), as well as the inflection class in case of base stems. Prefix entries specify the part of speech and origin feature of their derivational basis. Suffix entries encode the part of speech, stem type (derivation or compounding), and origin of the derivational basis. Additionally, they encode the part of speech, stem type, origin, word form class (simplex, prefix derivation, suffix derivation), and inflection class (in case of a base stem) of the word form resulting from the suffix derivation. These agreement features are used to filter out incorrect combinations of stems, affixes and inflectional endings. Morpho-phonological rules (see (Schiller, 1996) for more details) are applied to generate the correct surface forms. The tagset contains 572 tags.

A typical output of SMOR (for a complete overview, see (Schmid et al., 2004)) encodes the properties of the morphemes that form part of the word and the morphological processes applied:

- Part of Speech: <V>, <ADJ>, <NN>, <KSF>[3], ...
- Affixation: <PREF>, <SUFF>, ...
- Information on orthography <OLDORTH> (old orthography rules of German), <NEWORTH>

Example (1), repeated from example (2) in (Schmid et al., 2004), demonstrates the SMOR analysis of the adjective *unübersetzbarstes* ("most untranslatable"). The adjective is derived from the verb *übersetzen* ("[to] translate"). The suffix *unübersetz·**bar**·stes* derives an adjective (<+ADJ>) which is marked with a "+" to indicate the part of speech of the resulting word form. As it is a superlative form (*unübersetzbar·**st**·es*), it is marked by <Sup>. The adjective is negated with a respective prefix (***un***·*übersetzbarstes*) and it ends in a specific suffix (*unübersetzbarst·**es***), marking a neuter nominative singular form.

(1) un<PREF>übersetzen<V>bar<SUFF><+ADJ>
<Sup><Neut><Nom><Sg>

We have implemented a webservice based on SMOR which takes a list of word forms as input and returns their morphological analyses. The SMOR webservice is part of the WEBLICHT webservice infrastructure (Hinrichs et al., 2010). The webservice input and output are encoded in XML following the TCF standard (Heid et al., 2010). WebLicht is currently in an experimental stage and will become publicly available in the near future.

2. Tasks in morphology evaluation

2.1. Past experience

A number of morphology evaluation projects have been accomplished, from the MORPHOLYMPICS of 1994 (Hausser, 1996) to today's annual MORPHO CHALLENGES (Kurimo and Varjokallio, 2008). In both activities, Gold Standards have been provided by the organizers of the evaluation campaigns, and the systems submitted were evaluated by comparing their analysis results to these standards. The Gold Standards themselves were produced manually or semi-automatically, and issues of their design and creation were hardly discussed in the respective publications. While MORPHOLYMPICS required the analysis of both, complete texts and lists of word forms, MORPHO CHALLENGE projects provide target analyses for a given word list. The contestants (analysers capable of unsupervised learning) then have to deliver exactly these analyses.

2.2. Features to be examined

An increasing number of linguistic services for end users can be found on the web or is in preparation, e.g. WEBLICHT[4] or LANGUAGE GRID[5]. Because users should be capable of deciding which service best fits their needs, they would expect that each of the offered tools is validated in a way comparable to the others. Concerning morphological analysers, evaluation results on the following issues should be readily available:

- Coverage (with respect to corpus data):
  - Lexicon: lemmas, assignment of POS;

[1] D-SPIN is the German counterpart of the European Research Infrastructure project CLARIN (`www.clarin.eu`). It is supported by the Bundesministerium für Bildung und Forschung (BMBF-FKZ: 01UG0801A, 2008-2010).

[2] Additional compounding and derivation stems are generated automatically; for the notion of simplex versus compounding versus derivation stems, see (Fuhrhop, 1998, Fitschen, 2004).

[3] KSF stands for **K**ompositions**S**tamm**F**orm, and marks a stem used for compounding only.

[4] WEBLICHT weblicht.sfs.uni-tuebingen.de/englisch/weblicht

[5] GRID: langrid.nict.go.jp

  - Rules: inflection, word formation (derivation/compounding)

- Assessment of the granularity of the analyses (especially for word formation):
lists of morphemes versus word formation "history" versus word structure.

The word forms contained in the Gold Standard should be **near-representative**, and the evaluated system should correctly and exhaustively analyse them in terms of lemmatisation and POS assignment.
The Gold Standard has therefore to contain a large number of word forms, in order to (i) cover all relevant inflectional classes of a language, and to (ii) contain a sufficient number of medium to high frequency homographs. Such homographs are not systematic and appear just in single word forms: *sichere* is *sichern*$_{\mathrm{V}}$ + 1$^{\mathrm{st}}$ person singular of *sichern* versus *sicher*$_{\mathrm{ADJ}}$ + {fem. sg. | any gender plural}, but *sicherem*$^{\mathrm{ADJ}}$ (masc. dativ sg. of *sicher*$_{\mathrm{ADJ}}$) is not homographuous with any other form. Furthermore, syncretistic forms should be covered in full. Feminine nouns ending in *-en*, for example, must be analysed as nominative, dative, accusative and genitive plural.
A basic issue for word formation is to decide upon the level of analysis. Morpheme sequences, as offered by SMOR or by GERTWOL[6] as a default, are e.g. not very useful for the improvement of statistical machine translation: For German-to-English, for example, compound splitting would avoid one-to-many alignments (e.g. *Antrags*$_1$·*einreichung*$_2$ – "submission$_2$ of (the) proposal$_1$"). Compound splitting alone may therefore be more useful here, and may be a level on which to evaluate a system.
On the other hand, a full morphological structure tree (as offered e.g. by the CANOO[7] tools) is useful for morphosemantic analysis.
Particular difficulties arise from the combination of derivation and compounding in German, and, again, from occasional (unwanted) homography: many morphological systems tend to over-generate. For example, the word *Abteilungen* ("departments") may receive, in addition to the correct analysis as an *-ung*-normalisation of *abteilen* ("[to] divide"), a spurious analysis as a compound of several nouns, i.e. *Abtei*$_{\mathrm{NN}}$ ("abbey") and the plural form of *Lunge*$_{\mathrm{NN}}$ ("lung"). On the other hand, N+N homography also occurs: *Staubecken* is indeed to be split in two ways: *Stau·becken* ("reservoir") and *Staub·ecken* ("dusty corners").
The evaluation results (expected by the Gold Standard versus analyses provided by the tool) should be given in the following metrics (in terms of precision and recall) to make them comparable:

- precision: correct versus incorrect analyses;

- recall: existing versus missing analyses.

Concerning the calculation of precision and recall, any metrics applied is to take all expected analyses for each word form into account and, by comparing the output of the tool with them, sorted each into one the four categories: true/false positive and true/false negative:

- true positive: the candidate is appropriate and the analysis is correct;

- false positive: the candidate is appropriate and the analysis is incorrect or: the candidate is not appropriate, but the tool does deliver an analysis;

- true negative: the candidate is not appropriate and the tool does not provide an analysis

- false negative: the candidate is appropriate, but the analysis/analyses are not provided

3. Methodology

3.1. Creating a Gold Standard

Generally two requirements can be stated for the design of a near-representative Gold Standard when random text input is to be expected: it must have a good coverage in terms of frequent word forms, but also of derivational patterns and inflectional paradigms of the respective language. Highly frequent word forms only pose a challenge for the lexicon, but not for the rule component of a morphology system because they usually are not morphologically complex. Therefore, not only a list of frequent word forms is of interest, but also word forms of medium frequency have to be included.
Secondly, depending on what users may require in terms of granularity of analysis, we expect at least two versions to be made available: One describing words on the basis of morpheme sequences, e.g. listing the parts of compounds and containing derivation and inflectional information, and the other containing inflectional information only (on the basis of the lexicalised word). Other versions of the Gold Standard containing only compounding information or containing structural information in addition may follow.

3.2. Utilizing SMOR to design the Gold Standard

Systematic testing of a morphological system may begin with a list of medium-to-high-frequency word forms, extracted from corpora and sorted by their POS. Words which have the same inflectional paradigm can be categorised according to their inflectional paradigms such that each paradigm is tested.
For complex words, it is to be decided for each potential user group whether to prefer an analysis of the whole word as a lexicalised form, or whether it should be decomposed according to supposedly productive rules. We use SMOR results as a base for creating the Gold Standard, its results therefore must be thoroughly examined and enhanced. The SMOR analysis results of *Möglichkeit* ("possibility") in table 1, for example, show that the full version of SMOR delivers two analyses with different granularity at once, cf. 1 – 4 and 9 – 12. Analyses 5 – 8 demonstrate that SMOR over-generates in allowing any verb stem (*mögen* = [to] like) to merge with the adjectival suffix *-lich*, while *möglich* is synchronously not perceived as complex. Concerning this word form, we will select analyses 9 – 12 for the fine grained version of the Gold Standard.

[6] Gertwol: www2.lingsoft.fi/cgi-bin/gertwol
[7] Canoo: www.canoo.net

1	Möglichkeit⟨+NN⟩⟨Fem⟩⟨Dat⟩⟨Sg⟩
2	Möglichkeit⟨+NN⟩⟨Fem⟩⟨Acc⟩⟨Sg⟩
3	Möglichkeit⟨+NN⟩⟨Fem⟩⟨Gen⟩⟨Sg⟩
4	Möglichkeit⟨+NN⟩⟨Fem⟩⟨Nom⟩⟨Sg⟩
5	mögen⟨V⟩lich⟨ADJ⟩⟨SUFF⟩keit⟨SUFF⟩⟨+NN⟩⟨Fem⟩⟨Dat⟩⟨Sg⟩
6	mögen⟨V⟩lich⟨ADJ⟩⟨SUFF⟩keit⟨SUFF⟩⟨+NN⟩⟨Fem⟩⟨Acc⟩⟨Sg⟩
7	mögen⟨V⟩lich⟨ADJ⟩⟨SUFF⟩keit⟨SUFF⟩⟨+NN⟩⟨Fem⟩⟨Gen⟩⟨Sg⟩
8	mögen⟨V⟩lich⟨ADJ⟩⟨SUFF⟩keit⟨SUFF⟩⟨+NN⟩⟨Fem⟩⟨Nom⟩⟨Sg⟩
9	möglich⟨ADJ⟩keit⟨SUFF⟩⟨+NN⟩⟨Fem⟩⟨Dat⟩⟨Sg⟩
10	möglich⟨ADJ⟩keit⟨SUFF⟩⟨+NN⟩⟨Fem⟩⟨Acc⟩⟨Sg⟩
11	möglich⟨ADJ⟩keit⟨SUFF⟩⟨+NN⟩⟨Fem⟩⟨Gen⟩⟨Sg⟩
12	möglich⟨ADJ⟩keit⟨SUFF⟩⟨+NN⟩⟨Fem⟩⟨Nom⟩⟨Sg⟩

Table 1: Current SMOR analysis of *Möglichkeit* ("possibility")

3.3. Testing in two "stages"

When processing 3,000 word forms with SMOR, several 10,000 analyses are produced. To save time, we opt for examining SMOR results in two "stages": After the word form lists have been created automatically, they are each processed with SMOR. The word forms that are not identified by SMOR are then collected (cf. table 2). Of the remaining forms, only the simplest[8] analyses are selected for a first manual categorisation (plausible, not plausible and unclear). By getting this first overview, we plan to find problem categories, some of which may possibly be identified semi-automatically. During the second stage, implausible analyses that were collected in the first step, are deleted automatically. We thereafter examine all remaining analyses in detail. In parallel, we generate the Gold Standard, partially by selecting correct analyses of the intended granularity, partially by changing the granularity of analyses provided by SMOR, and partially by suggesting additional analyses.

4. Preliminary results

We are in the process of preparing the fine grained version of the Gold Standard semi-automatically on the basis of frequent word forms. Currently, we prepare Gold analyses of about 1,000 words per productive word class, and categorise them according to compounding, derivational and inflectional patterns. The task is automated wherever possible.

An interesting question coming up during the process is that of finding guidelines stating which granularity of analysis is to be achieved: For example, two morphological analysers (SMOR, CANOO) analyse the German noun *Zukunft* ("future") as being derived from the verb *zu·kommen* ("[to] approach"), though this analysis is semantically rather opaque for a modern speaker of the language. It appears that the process works similarly for the noun *Ankunft* ("arrival") which is correctly derived from the verb *an·kommen*. The word part *·kunft* is developed from the Middle High German word *kumft*; an adverb/adjective *künftig* (Middle High German *kumftig*) derived from *kumft* is still in use. We have to ask ourselves if a speaker of German today will accept an analysis of *Zukunft* being derived from *zu·kommen*? The systems are supposed to be descriptive, not etymological.

Another issue is that of developing guidelines for accepting and marking neoclassical stems. So far, SMOR analyses *Akademikerin* as shown in (2-a). This formative *Akadem*, as all neoclassical stems, is marked with the POS of the greek or latin word it was derived from[9]. Therefore, SMOR analyses like (2-a) and (2-d) are based on sound linguistic arguments. On the other hand, neoclassical stems only allow special compounding and derivational processes. For a Gold Standard, we therefore suggest to mark stems like, e.g. *Akadem*$_{\mathrm{NN}}$, or *informat*$_{\mathrm{V}}$ with an additional information, e.g. *nc* (alternative, but less plausible analyses could be (2-c) and (2-f)), leading to the analyses (2-b) and (2-e). This issue, however, is still under discussion.

(2) a. Akadem<NN>ik<NN><SUFF>er<NN><SUFF>
in<SUFF><+NN>
b. Akadem<ncNN>ik<NN><SUFF>er<NN><SUFF>
in<SUFF><+NN>
c. Akadem<KSF>ik<NN><SUFF>er<NN><SUFF>
in<SUFF><+NN>
d. informat<VV><SUFF>ion<SUFF><+NN>
e. informat<ncVV><SUFF>ion<SUFF><+NN>
f. informieren<VV><SUFF>ion<SUFF><+NN>

The acceptance of other, less opaque derivational analyses, like, e.g. whether to accept analyses where nouns like *Betrieb* ("company") are being derived from *betreiben* ("[to] operate/run (e.g. a company)") can be decided upon alongside more general guidelines, cf. paragraph 4.2.2.

4.1. Creating word form lists

The 1.3 billion token DeWaC corpus of German web texts (Baroni and Kilgarriff, 2006) serves as our data source. The original encoding of the corpus is mixed (latin-1, utf-8 and others). Being a web corpus, the data is noisy and contains a fair amount of duplicates, typographic errors and nonwords. We first changed its encoding to latin-1 to enable a stable processing with scripting tools. The sentences of the corpus were then sorted uniquely while taking their source information into account, i.e. double sentences do still occur if they appear on different URLs. After some basic, sentence-wise cleaning done similarly to the procedures described in (Quasthoff et al., 2006), we parsed the remaining

[8] The simplest analysis of a set is the one with the smallest number of derivation and compounding steps.

[9] Concerning neoclassical forms, see also (Luedeling et al., 2001)

sentences with the IMS internal dependency parser FSPAR (Schiehlen, 2003). FSPAR reports the nodes that it cannot attach to the sentence parse tree; thus we are able to calculate a rough *error rate* per sentence in order to estimate parse quality. Only sentences with a low error rate were used for further processing. The resulting subcorpus is called SDEWAC and contains ca. 880 million tokens of a wide variety of domains. It is tokenised with an IMS internal tokeniser (Schmid, 2000) and tagged with the Tree-Tagger (Schmid, 1994).

Thereafter, a python script automatically extracted lists of word forms together with information on their frequency for the word classes (common) nouns, (finite main) verbs and adjectives[10]. The resulting lists contained 1.985.291 noun candidates[11], 142.701 (finite main) verb candidates, and 303.197 adjective candidates.

As said above, a Gold Standard should contain word forms of medium to high frequency of occurrence. Calculating the median of word frequencies of the corpus would result in a number near to 1, because most word forms in (any) corpus only occur once or twice (hapax and dis legomina)[12]. Hence we arbitrarily opted for calculating a median of the frequencies for each word class after the word forms that only occurred once or twice had been deleted from the list. The resulting median values were 6 for nouns, 11 for verbs, and 8 for adjectives. Using a python script, we thereafter randomly selected 1,000 word forms per category out of the frequencies ranging between median and highest.

4.2. Results of the first stage

During the first stage, we checked the words that SMOR did not recognise (out of the list of 1,000 word forms of each word class): these summed up to 232 nouns, 151 verbs and 177 adjectives. While most of these word forms had either been written incorrectly (typos) or were foreign language material (true negatives), a number of correctly spelled and tagged word forms were also contained in the lists. However, most of them belonged to specific terminology (medicine, geography, etc.). Results and examples are shown in table 2.

word class	**total**	**true neg**	**false neg**	**examples**
NN	232	217	15	Terpenoide
VVFIN	151	147	4	infundiert
ADJA	177	143	34	austriakischen

Table 2: Word forms not identified by SMOR

Concerning the remaining word forms, we examined the 2,989 simplest analyses of 768 nouns, the 4,045 simplest analyses of 849 verbs, and the 6,011 simplest analyses of 823 adjectives. All of these analyses were categorised into "possible" (true positives), "false" (false positives) or "undecided" to indicate that a decision on their correctness was postponed to the second stage. In paragraph 4.2.2., we will show categories of some error/problem cases. Note that results and cases of homography of past participle (pp) verb and adjective (e.g. *beseitigt* ("eliminated")) are counted separately in table 3 which shows the results.

word class	**no. of**	**simplest ana.**	**true pos.**	**false pos.**	**rest** (pp)
NN	768	2,989	2,580	238	171
%		100	86.3	8.0	5.7
VVFIN	849	4,045	2,854	59	143 (989)
%		100	70.55	1.5	3.5 (24.4)
ADJA	823	6,011	5,159	507	174 (171)
%		100	85.8	8.4	2.9 (2.8)

Table 3: Word forms identified by SMOR - a first examination of the simplest analyses

Some wrongly spelled words were found during this processing step, e.g. the adjective *erbschaftsteuerrechliche* is analysed by SMOR though the word form does not exist (the word form is correctly spelled as *erbschaftssteuerrechtliche*, an adjective related to "laws on death duty"). These analyses are counted as false positives. Others had been wrongly tagged, e.g. *kalken* ("[to] coat with lime") was found in the noun list, though it constitutes (the infinitive) of a verb. Such word forms are counted as true negatives because they occur in the wrong list.

4.2.1. Second stage: An intermediate evaluation result of interpreting analyses of 100 nouns

The second stage of processing is necessary for the creation of the foreseen fine grained version of the Gold Standard and to produce a detailed documentation of SMOR. In this step, we examine all analyses produced by SMOR for the identified word forms. Current work concentrates on the list of nouns.

An intermediate result on 100 nouns can be reported here. SMOR generated 826 analyses for these word forms, 586 of which are counted as true positives of different granularity, 221 are seen as false positives and 9 as true negatives. We found that 26 analyses were missing, these were counted as false negatives. Summing up to 852 analyses of the 100 nouns, SMOR performed with a precision of 0.72 and a recall of 0.96 (F-measure = 0.82).

4.2.2. Preparing the fine grained version of the Gold Standard

As mentioned in the paragraphs above, there are several possibilities when deciding on the granularity of of derivational and compositional analyses. Some cases must be decided for individual words, others can be generalized. To demonstrate this issue, we come back to the German compound noun *Priesterweihe*, which was already mentioned

[10] We only selected attributive adjectives (tagged ADJA; STTS tagset, cf. (Schiller et al., 1995))

[11] Schmid's Tree-Tagger (Schmid, 1994) has a reported accuracy of more than 98% on clean text. Web corpora can be assumed to be more noisy than others, e.g. news paper corpora, hence we expect the accuracy of any tagger to be lower than usual. Therefore, a number of "non-words" and words with wrongly assigned parts of speech were expected to appear in the word lists. These present true negatives in our accuracy tests.

[12] The phenomenon was described by Zipf in 1949, cf. (Manning and Schütze, 2001, pp. 23,35)

in paragraph 1.1. The analyses in (3) reflect the two nouns *Priester* ("priest") and *Weihe* ("ordination"). We could leave the analysis like that, cf. (3-a). However, we can also see the latter part of the compound as a noun derived from the verb *weihen* ("[to] ordinate"), as done (by SMOR[13]) in (3-b).

(3) a. Priester<NN>Weihe<+NN>
b. Priester<NN>weihen<V><SUFF><+NN>

Different users may want a different granularity of analysis. For our fine grained version of the Gold Standard, we follow the "as fine as it gets"-strategy and we therefore opt for the latter analysis (3-b), i.e. we include derivational analyses whenever possible.
Secondly, we handle nominalisations with ablaut exactly as those without: a noun like, e.g. *Betrieb* is seen as being derived from the verb *betreiben*, as shown in (4) and (5), where the word forms *Betriebsstörung* and *Speicherbedarf* are analysed. Such analyses are seen as being structurally identical with the analysis of e.g. *Beleg* in (6) where the nominalisation is without ablaut.

(4) be<VPREF>treiben<V>stören<V>ung<SUFF><+NN>

(5) Speicher<NN>be<VPREF>dürfen<V><SUFF><+NN>

(6) be<VPREF>legen<V><SUFF><+NN>

In the course of our work, we detected problems with the rule component of SMOR (cf. table 4 showing the analyses of *Abschlussklausur* ("course examination")): as a result of several rules being processed in parallel, identical derivational/compounding analyses are produced twice, each with their elements in a different order. Superfluous analyses like 5–8 of table 4 were not counted as being false positives, but marked with a note "rule" do indicate a respective error.
Lastly, we found the rule component of SMOR to produce other superfluous analyses with compounding stems, mixing up upper and lower case, cf. table 5 showing the analyses of *Agrarbetriebe* ("agricultural farms"). Note that none of these analyses will be taken for the Gold Standard, as we suggest *Betrieb* to be further analysed as a nominalisation of the verb *betreiben*, cf. paragraph 4.2.2.
For the first 100 nouns of our list (which contains 768 nouns in total), we selected 270 true positive analyses in total. In addition to 26 missing ones, we suggest to modify 53 of the true positives. 20 analyses were identified as being superfluous and were reported for rule correction. The missing analyses of 15 word forms not analysed at all (cf. table 2) will be added later.

4.3. Preparing a version of the Gold Standard describing inflectional information only

SMOR is capable of providing information on inflection only by simplifying fine grained analyses; the resulting version of the Gold Standard does hence not have to be evaluated explicitly. Rows 1 – 4 in table 6 reflect the simplified output for the example of *Priesterweihe* ("Priest ordination"). Rows 5 – 8 demonstrates the respective fine grained output for comparison reasons.

5. Summary and future work

This paper attempts to contribute to evaluation methodology in terms of the design and creation of a Gold Standard for morphology systems. We have described a number of issues to be considered from the developer's and the users' perspective and applied them to an ongoing evaluation of a morphological system, SMOR(Schmid et al., 2004).
To satisfy the needs of potential users, the following aspects of a morphological system need to be documented:

- The tools' knowledge resources: provision of information about the size and the content of the lexicons
- The test lists: provision of the versions of the Gold Standard (in different grades of granularity)
- The way in which these versions were produced
- The tools' results when processing the test list in terms of recall and precision

We are in the process of designing a fine grained Gold Standard that may be utilized for evaluating morphological analysers of German. A simplified version providing only inflectional information is also in preparation. This project is planned to be finished in spring 2011. In parallel, we comprehensively test and document features of the analyser SMOR which will be offered as a web service in the near future.

6. Acknowledgements

We are very grateful to the following persons: Fabienne Fritzinger, not only for converting the source corpus, but also for offering valuable advice on the methodology for a evaluation of SMOR; Kerstin Eckart, who supported us concerning a fast method to judge parsing results of FSPAR; and Alexandra Kolb and Ronny Jauch for generating word lists, processing them with SMOR and manually checking and commenting on its results.

7. References

Christina Bankhardt. 2009. D-SPIN - Eine Infrastruktur für Deutsche Sprachressourcen. *Sprachreport*, 25(1):30 – 31.

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*, pages 87 – 90, Trento, Italy. EACL.

EAGLES. 1996. *Evaluation of Natural Language Processing Systems, EAG-EWG-PR.2*. EAGLES, final report edition, October.

Arne Fitschen. 2004. *Ein Computerlinguistisches Lexikon als komplexes System*, volume 10 no. 3 of *AIMS – Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung*. Lehrstuhl für Computerlinguistik, Universität Stuttgart, Stuttgart.

Nanna Fuhrhop. 1998. *Grenzfälle morphologischer Einheiten*. Number 57 in Studien zur deutschen Grammatik. Stauffenburg, Tübingen.

[13] In examples (3), inflectional information produced by SMOR is deleted for demonstration reasons.

1	Abschluss<NEWORTH><NN>Klausur<+NN><Fem><Nom><Sg>
2	Abschluss<NEWORTH><NN>Klausur<+NN><Fem><Gen><Sg>
3	Abschluss<NEWORTH><NN>Klausur<+NN><Fem><Acc><Sg>
4	Abschluss<NEWORTH><NN>Klausur<+NN><Fem><Dat><Sg>
5	Abschluss<NN><NEWORTH>Klausur<+NN><Fem><Nom><Sg>
6	Abschluss<NN><NEWORTH>Klausur<+NN><Fem><Gen><Sg>
7	Abschluss<NN><NEWORTH>Klausur<+NN><Fem><Acc><Sg>
8	Abschluss<NN><NEWORTH>Klausur<+NN><Fem><Dat><Sg>

Table 4: SMOR analyses of `Abschlussklausur` (new orthography)

1	Agrar<KSF>Betrieb<+NN><Masc><Dat><Sg><Old>
2	Agrar<KSF>Betrieb<+NN><Masc><Nom><Pl>
3	Agrar<KSF>Betrieb<+NN><Masc><Gen><Pl>
4	Agrar<KSF>Betrieb<+NN><Masc><Acc><Pl>
5	agrar<KSF>Betrieb<+NN><Masc><Dat><Sg><Old>
6	agrar<KSF>Betrieb<+NN><Masc><Nom><Pl>
7	agrar<KSF>Betrieb<+NN><Masc><Gen><Pl>
8	agrar<KSF>Betrieb<+NN><Masc><Acc><Pl>

Table 5: SMOR analyses of `Agrarbetriebe`

	word form	analysis	lemma
1	Priesterweihe	NN.Fem.Nom.Sg	Priesterweihe
2	Priesterweihe	NN.Fem.Gen.Sg	Priesterweihe
3	Priesterweihe	NN.Fem.Acc.Sg	Priesterweihe
4	Priesterweihe	NN.Fem.Dat.Sg	Priesterweihe
	word form	analysis	
5	Priesterweihe	Priester<NN>weihen<V><SUFF><+NN><Fem><Dat><Sg>	
6	Priesterweihe	Priester<NN>weihen<V><SUFF><+NN><Fem><Gen><Sg>	
7	Priesterweihe	Priester<NN>weihen<V><SUFF><+NN><Fem><Acc><Sg>	
8	Priesterweihe	Priester<NN>weihen<V><SUFF><+NN><Fem><Dat><Sg>	

Table 6: Comparison of inflectional and fine grained versions of the Gold Standard

Roland Hausser, editor. 1996. *Linguistische Verifikation. Dokumentation zur Ersten Morpholympics 1994.* Niemeyer, Tübingen, Germany.

Ulrich Heid, Helmut Schmid, Kerstin Eckart, and Erhard Hinrichs. 2010. A corpus representation format for linguistic web services: the D-SPIN Text Corpus Format and its relationship with ISO standards. In *Proceedings of the seventh international conference on Language Resources and Evaluation (LREC)*, Valetta, Malta, 19 – 21 May 2010. European Language Resources Association (ELRA).

Marie Hinrichs, Thomas Zastrow, and Erhard Hinrichs. 2010. WebLicht: Web-based LRT Services in a Distributed eScience Infrastructure. In *Proceedings of the seventh international conference on Language Resources and Evaluation (LREC)*, Valetta, Malta, 19 – 21 May 2010. European Language Resources Association (ELRA).

Mikko Kurimo and Matti Varjokallio. 2008. Unsupervised morpheme analysis evaluation by a comparison to a linguistic gold standard – morpho challenge 2008. In *Working Notes for the CLEF 2008 Workshop*.

Anke Luedeling, Tanja Schmid, and Sawwas Kiokpasoglou. 2001. On neoclassical word formation in german. In Geert Booij and Jaap van Marle, editors, *Yearbook of Morphology*, pages 253 – 283. Kluwer, Dordrecht.

Christopher D. Manning and Hinrich Schütze. 2001. *Foundations of Statistical Natural Language Processing*. MIT Press, Boston, MA, 4 edition.

Uwe Quasthoff, Matthias Richter, and Christian Biemann. 2006. Corpus portal for search in monolingual corpora. In *Proceedings of 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1799 – 1802, Genoa, Italy.

Michael Schiehlen. 2003. A Cascaded Finite-State Parser for German. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003)*, pages 163 – 166, Budapest, Hungary.

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. 1995. Vorläufige Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical report, Universität Stuttgart, Institut für maschinelle

Sprachverarbeitung, and Seminar für Sprachwissenschaft, Universität Tübingen.

Anne Schiller. 1996. Deutsche flexions- und kompositionsmorphologie mit PC-KIMMO. In Roland Hausser, editor, *Linguistische Verifikation. Dokumentation zur Ersten Morpholympics 1994*, Tübingen. Niemeyer.

Tanja Schmid, Anke Lüdeling, Bettina Säuberlich, Ulrich Heid, and Bernd Möbius. 2001. Deko – ein system zur analyse komplexer wrter. In Henning Lobin, editor, *Proceedings der GLDV-Frühjahrstagung 2001*, pages 49 – 57, Universitt Gießen, 28. – 30. Mrz.

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. A German Computational Morphology Covering Derivation, Composition, and Inflection. In *Proceedings of the IVth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 1263 – 1266, Lisbon, Portugal.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pages 44 – 49, Manchester, UK.

Helmut Schmid. 2000. Unsupervised Learning od Period Disambiguation for Tokenisation. `http://www.ims.uni-stuttgart.de/˜schmid`. Internal Report, IMS, University of Stuttgart.

Helmut Schmid. 2005. A programming language for finite state transducers. In *Proceedings of the 5th International Workshop on Finite State Methods in Natural Language Processing (FSMNLP 2005)*, Helsinki, Finland.

Nancy L. Underwood. 1998. Issues in Designing a Flexible Validation Methodology for NLP Lexica. In A. Rubio, N Gallardo, R. Castro, and A. Tejada, editors, *Proceedings of the First International Conference on Language Resources and Evaluation*, volume 1, pages 129 – 134, Granada, Spain, 28 – 30 May.