Reconstructing Complete Lemmas for Incomplete German Compounds

Anonymous

No Institute Given

Abstract. This paper discusses elliptical compounds, which are frequently used in German in order to avoid repetitions. This phenomenon involves truncated words, mostly truncated compounds. These words pose a challenge in POS tagging and lemmatization, which often leads to unknown or incomplete lemmas. We present an approach to reconstruct complete lemmas of truncated compounds in order to improve subsequent language technology or corpus linguistic applications. Results show an f-measure of 95.6% for the detection of elliptical compound patterns and 86.4% for the correction of compound lemmas.

1 Introduction

Many languages use elliptical constructions in order to avoid repetitions and to allow concise wordings. In this paper we focus on elliptical compounds in German coordination constructions. We use the term "elliptical compound"[1] to refer to truncated words like *Schnee-* in coordinated constructions such as *Schnee- und Lawinenforschung* (snow [research] and avalanche research). These truncated words typically end with a hyphen, which stands for the last part of a full compound following the conjunction. The full coordination in the example would be *Schneeforschung und Lawinenforschung*.

Since we want to have access to all compounds in our corpus in a unified way, we are interested in resolving the hyphen references of truncated words. Current Part-of-Speech (PoS) taggers for German usually assign the dummy tag TRUNC to such elliptical compounds (Thielen et al., 1999). Most current lemmatizers work without regard to context and therefore they cannot assign the full compound as lemma for a truncated word. They either use the word form as lemma, or they opt for the dummy lemma "unknown". Since incomplete lemmas are a stumbling block for any type of machine processing, our goal is to overcome this restriction and to deliver a full lemma for a truncated word based on an analysis of the coordination construction.

Towards this goal we have collected the most typical coordination patterns in German that involve elliptical compounds. We have developed a program that, when triggered by a truncated word, determines the coordination pattern, splits the full compound of the construction into its elements, and uses the

[1] We are aware of the fact that not all the cases matching our patterns are compounds. However, compounds are by far the most frequent and typical.

last segment of this full compound to generate the full lemma of the truncated word. The freedom in creating elliptical compounds in German and the resulting diversity of constructions turn this into an interesting challenge.

We have developed and tested our lemma reconstructor for the German part of the Text+Berg corpus, a large collection of texts from the Swiss Alpine Club. However, we believe that the coordination patterns and the methodology are applicable to other corpora as well.

In the next section, we briefly introduce some related work, and we describe the linguistic phenomenon of elliptical compounds. In section 3 we introduce the Text+Berg corpus and its characteristics. In section 4, we give an overview of the general system architecture, illustrated with examples. Section 6 presents our evaluation of the lemma reconstructor and mentions some problematic cases.

2 Elliptical Compounds

The linguistic properties of compounds are widely studied and some of these works include sidesteps on elliptical compounds with hyphens. There exist several types of hyphens; one is used to break single words into parts if the word continues in the following line, another to join separate words into one word. For our work, the third usage is important; the suspended hyphen which marks the truncated word (Bredel, 2008). Srinivasan (1993) sees the suspended hyphen as a morpheme placeholder whereas he describes the other two usages as “word breaks” since they break connected chains of letters.

According to the official spelling rules 98[2] and the Duden dictionary rule 31[3] a suspended hyphen can replace the lexeme which the compounds have in common. This can be the first and/or the last lexeme:

- first lexeme: *Bergkameradschaft und -hilfe* (moutain fellowship and [mountain] help)
- last lexeme: *laut- und spurlos* (sound[less] and traceless)
- first and last lexeme: *Sonnenauf- und -untergang* (sun[rise] and [sun]set)

The truncated word is described by Eisenberg (2004) as a separate syntactic base form; the “word rest”. This indicates that something has been omitted, which, however, can be regained in the surroundings.

The first challenge is to find all the patterns of such compounds, the second is the correct word segmentation. The latter task is complicated by the freedom of merging words as a common type of word formation in German. This is also shown by some word formations occurring in the Text+Berg corpus (**?**)[4] e.g. *Edelweissromantik* (Edelweiss romance), *Akrobatentänzerverein* (acrobat dancer club) or *Wegwerflandschaften* (disposable landscapes).

In order to obtain an overview of the frequency of the elliptical compound phenomenon we investigated the TIGER treebank (Brants et al., 2002), a collection of 50,474 German syntax trees corresponding to 888,299 tokens. In this

[2] canoo.net/services/GermanSpelling/Amtlich/Inter-punktion/pgf98.html, 2.10.2012

[3] www.duden.de/sprachwissen/rechtschreibregeln/binde-strich#K31, 2.10.2012

[4] Some references are suppressed for anonymous reviewing.

corpus we found 1226 sentences (about 2% of all sentences) with a total of 1362 tokens which are tagged as TRUNCated words. Figure 1 shows an example with two such truncated words *Natur-, Umwelt-*. Note the reduced lemmas for these words. Since the full compound *Lebensschutz* is not given with segmentation boundaries, it is far from trivial to reconstruct the complete lemmas for the truncated words.

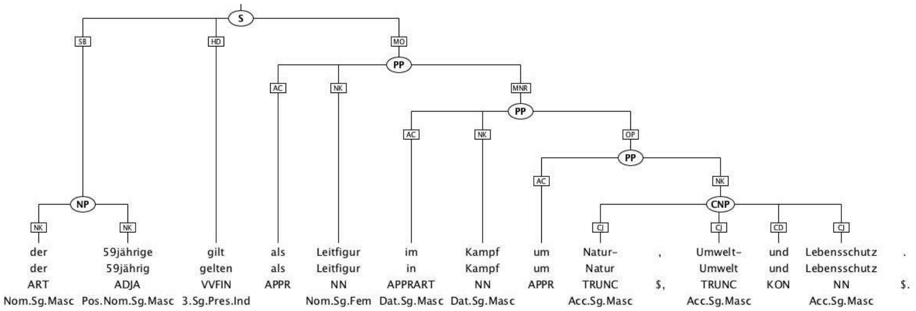

S
SB
HD
MO
PP
AC
NK
MNR
PP
AC
NK
OP
PP
AC
NK
CNP
CJ
CJ
CD
CJ
NP
NK
NK
der der ART Nom.Sg.Masc
59jährige 59jährig ADJA Pos.Nom.Sg.Masc
gilt gelten VVFIN 3.Sg.Pres.Ind
als als APPR
Leitfigur Leitfigur NN Nom.Sg.Fem
im in APPRART Dat.Sg.Masc
Kampf Kampf NN Dat.Sg.Masc
um um APPR
Natur– Natur TRUNC Acc.Sg.Masc
, $,
Umwelt– Umwelt TRUNC Acc.Sg.Masc
und und KON
Lebensschutz Lebensschutz NN Acc.Sg.Masc
. $.

Fig. 1. Example sentence from the TIGER treebank with two truncated words

The statistics show that the TIGER treebank includes

- 6 TRUNC tokens with digits (e.g. *16- und 17jährigen* (16[-year-old] and 17-year-old))
- 110 TRUNC tokens that start with a lower case letter (e.g. *in- und ausländischen* (domestic and foreign), *bi- als auch multilaterale* (bi[lateral] and multilateral), *mittel- bis langfristig* (medium-[term] to long-term))
- 28 TRUNC tokens without hyphen whose classification is dubious (e.g. *des (ab/TRUNC) laufenden Jahrhunderts* (of the (expiring) continuous decade), *dem nachkolonialen/TRUNC und Nachkriegs-Vietnam* (the post colonial and post war Vietnam), *der Beteiligungs/TRUNC AG* (the holding corporation), *Fast/TRUNC Food*)

Words with initial hyphens are not tagged as TRUNC in the TIGER corpus. They are tagged as if the hyphens were not present, mostly they are regular nouns (NN).

- *Desintegrations-Ängste und -Erfahrungen/NN* (disintegration fears and [disintegration] experiences)
- *Bodenverwaltungs- und -verwertungsgesellschaft/NN* (land administration [corporation] and [land] usage corporation)
- *weder Arbeitslosengeld noch -hilfe/NN* (neither unemployment benefit nor [unemployment] aid)

In the TIGER corpus the lemmas of the truncated words correspond to the first lexeme of the compound. This means the hyphen is omitted, thus the truncated compound of *vier- bis fünfhundert* (four [hundred] to five hundred) is annotated with the lemma *vier* (four). If there is a linking element (*-s-* or *-es-*), it is also omitted which means that the truncated word of *Bundes- und Reichsbahn* (federal [railway] and state railway) is annotated with the lemma *Bund* (confederation).

The aforementioned examples show how much information gets lost, which implies that the completion of elliptical compounds is a useful step when annotating a corpus. With these reconstructions, a full text analysis on compound lemmas can be carried out which will improve the results of subsequent language technology tasks. We hypothesize that, for instance, machine translation systems can be improved with these completed lemmas. In any case, complete lemmas facilitate corpus searches in linguistic research.

Elliptical compounds are not restricted to German. They occur in similar ways in other compounding languages. For example, Swedish also uses the hyphen to mark truncated words in much the same way as German (e.g. *olje- och gasverksamheten* (oil and gas activities), *Nord- och Sydamerika*, *lång- och kortfristiga* (long and short-term)). The frequency of last lexeme truncations in the one million token Stockholm-Umeå corpus (Gustafson-Capková and Hartmann, 2006) is on the same order of magnitude as in the German TIGER corpus. In contrast, truncated words with hyphens for first lexeme omission are very rare in Swedish (e.g. *tillverkningsprocesser och -metoder* (production processes and methods)).

Our approach to reconstruct the lemmas of incomplete compounds also has a correspondence in non-compounding languages like English. Consider the coordination construction *oil and gas activities.* In order to interpret this coordination correctly, a system has to detect that both *oil* and *gas* are modifiers to *activities.* The resolution thus amounts to disambiguation of modifiers in complex coordinations.

3 The Text+Berg Corpus

We have digitized all yearbooks of the Swiss Alpine Club from 1864 until today. Most of the articles are in French and German, a few in English, Italian and Romansh. We scanned the texts and used OCR (optical character recognition) software to convert the scan images into text (**?**)[4]. We then structured the text by identifying article boundaries based on manually corrected tables of contents. We tokenized and split each text into sentences and determined the language of each sentence automatically. This procedure allows, amongst others, the recognition of German quotations in French articles and vice versa.

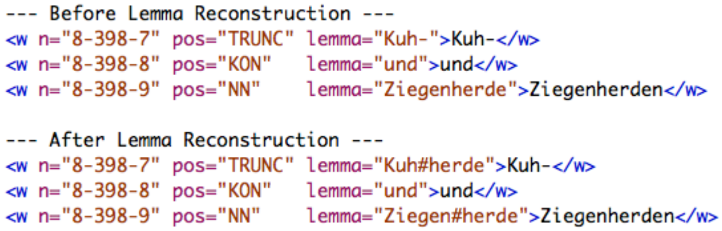
Tokenization of the corpus was a major undertaking because of the spelling idiosyncrasies over the 150 year time span and the different languages. In particular, apostrophes and hyphens caused problems. Hyphens that result from end-of-line word breaks were eliminated by checking whether the form without hyphen was clearly more frequent in the corpus than the form with the hyphen. Hyphenated compounds were left as one token (e.g. *SAC-Hütte*), but French combinations of verb + pronoun were split (e.g. *viennent-ils* → *viennent* + *-ils*) in order to facilitate PoS tagging. Word-final hyphens were left intact as they allow us to identify truncated words.

After tokenization the corpus was part-of-speech tagged and lemmatized with different parameter files for the TreeTagger (Schmid, 1994) for English, French, German and Italian. We added missing lemmas for German and French based on other dictionaries. Sentences in Romansh were not annotated since there is no PoS tagger and lemmatizer for this language. Subsequently special modules were run for named entity recognition (geographical names and person names) and for alignment between the French-German translated parts of the corpus. All the information is stored in XML files.

The Text+Berg corpus currently consists of 22.5 million tokens in German and 21.5 million tokens in French, out of which about 5 million tokens are translations (i.e. parallel texts). The Italian part has 0.3 million tokens, English and Romansh have less than 100,000 tokens each. The corpus is freely available for research purposes.

4 Architecture of our Lemma Reconstructor

Our Python script takes as input a corpus file in XML and two lists containing word frequencies and word segmentations. The input XML file represents a typical elliptical compound construction as depicted in figure 2. Each token is annotated with a unique identifier, a part-of-speech tag, and a lemma. After the reconstruction both the elliptical compound and the full compound have complete lemmas which include the segmentation boundary markers (see figure 2).



```
--- Before Lemma Reconstruction ---
<w n="8-398-7" pos="TRUNC" lemma="Kuh-">Kuh-</w>
<w n="8-398-8" pos="KON"   lemma="und">und</w>
<w n="8-398-9" pos="NN"    lemma="Ziegenherde">Ziegenherden</w>

--- After Lemma Reconstruction ---
<w n="8-398-7" pos="TRUNC" lemma="Kuh#herde">Kuh-</w>
<w n="8-398-8" pos="KON"   lemma="und">und</w>
<w n="8-398-9" pos="NN"    lemma="Ziegen#herde">Ziegenherden</w>
```

Fig. 2. XML example before and after our lemma reconstruction program

The task of completing lemmas can roughly be divided into two steps. On the one hand, our lemma reconstructor looks for patterns that contain elliptical compounds, on the other hand, found patterns are modified in order to complete the lemma of the truncated word with the missing part.

4.1 Pattern Matching

The main function runs through the parsed XML file looking for patterns. If a pattern matches, the task is forwarded to several modules in order to analyse the compound construction and identify the missing part. As the German part of the Text+Berg corpus is part-of-speech annotated with the Stuttgart Tübingen Tagset (Thielen et al., 1999), the patterns are defined using the same tags[5]. For some patterns more than one solution is possible. In such cases, all the possible solutions are generated and with the help of word frequencies the most likely solution is chosen. There are eleven specified patterns[6], some of which have several solution variants.

(1) TRUNC1 + $, + TRUNC2 + $, + TRUNC3 + KON + NN/NE/ADJA
⇒ TRUNC1 + WORD_L **&** TRUNC2 + WORD_L **&** TRUNC3 + WORD_L
→ *Kondordia-, Gleckstein-, Dossen- und Gauli__hütte__*
(Konkordia [Hut], Gleckstein [Hut], Dossen [Hut] and Gauli **Hut**)

(2) TRUNC1 + $, + TRUNC2 + KON + NN/NE/ADJA
⇒ TRUNC1 + WORD_L **&** TRUNC2 + WORD_L
→ *Wind-, Niederschlags- und Temperatur-__messungen__*
(wind [measurements], rain [measurements] and temperature **measurements**)

(3) TRUNC + KON + -WORD
⇒ TRUNC + -WORD_L **&** TRUNC_L + -WORD
→ *__Schiefer__schutt- und -platten__hang__* (**slate** scree [slope] and [slate] slab **slope**)

(4) TRUNC + KON + APPR + ART + NN/NE
(a) TRUNC + NN_L
→ *Nord- und auf der Ost__seite__* (north [side] and on the east **side**)
(b) TRUNC + APPR_L
→ *dies- und jen__seits__* (this [side] and on the other **side**)

(5) TRUNC + KON + ADJ/ADV/VVFIN/VVIZU/ VVPP/CARD/APPR + !NN/NE
⇒ TRUNC + WORD_L
→ *hinauf- und hinunter__geklettert__* ([climbed] up and **climbed** down)

(6) TRUNC + KON + ADJA/CARD/ADV + NN/NE
(a) TRUNC + NN
→ *Alpin- und sonstige __Rucksäcke__* (alpine [backpacks] and other **backpacks**)
(b) TRUNC + ADJ_L
→ *hilf- und gnaden__reichen__ Mutter* (help[ful] and merci**ful** mother)
(c) TRUNC + NN_L
→ *Sessel- und vier Ski__lifts__* (chair [lifts] and four ski **lifts**)

[5] In addition to the STTS tags ADJ stands for different adjectives, WORD for different parts of speech.

[6] / stands for *or*, ! for *not*, F/L for the first/last lexeme of a word, (a), (b), (c) for different solution variants, ⇒ implies that there is only one specified solution.

(7) TRUNC + KON + ART/APPR/APPRART + NN/NE
⇒ TRUNC + NN_L
→ *Fels- oder durch Schutt**massen*** (rock [material] or through scree **material**)

(8) TRUNC + KON + ART/CARD/ADV/APPR/APPR-ART + ADJA + NN/NE
(a) TRUNC + NN
→ *Languard- und die angrenzenden **Gebiete***
(Languard [areas] and the bordering **areas**)
(b) TRUNC + NN_L
→ *Walen- und dem Oberen Zürich**see***
([Lake] Walen and the Upper *Lake* Zurich)
(c) TRUNC + ADJA_L
→ *französisch- und der deutsch**sprachigen** Schweiz*
(French [speaking] and German **speaking** Switzerland)

(9) TRUNC + KON + NN
⇒ TRUNC + NN_L
→ *Eis- und Schnee**wänden*** (ice [walls] and snow **walls**)

(10) TRUNC + APPRART + NN/NE/ADJA
⇒ TRUNC + NN_L
→ *Nord- zum Süd**gipfel*** (north [peak] to the south **peak**)

(11) WORD1 + KON + -WORD2
⇒ WORD1_F + -WORD2
→ ***Berg**steiger und -führer* (**mountain** climber and [mountain] guide)

4.2 Compound Analysis, Solution Generation and Decision

If, for example, *Schnee- und Lawinenforschung* (snow [research] and avalanche research) has been found matching the pattern *(9) TRUNC + KON + NN*, the construction is split for further processing. First, the complete compound *Lawinenforschung* is analysed by Gertwol[7], a wide-coverage morphology system for German. Gertwol does de-compounding for all words where all segments are known to the system. It provides an analysis with four different segmentation symbols[8] to differentiate between different word-internal boundaries.[9] In our example, Gertwol delivers the segmentation *Lawine\n#forsch∼ung*. We disregard the linking element and the suffix segmentation, and so this compound consists of only two parts, i.e. *Lawine* and *forschung*. By recognizing the strong compound boundary (#), it is trivial to pick the second part for filling the lemma of the truncated word *Schnee-*.

[7] www.lingsoft.fi

[8] www2.lingsoft.fi/doc/gertwol/intro/segment.html, 2.10.2012

[9] # for a strong boundary, – for a weak boundary, \for a linking element, ∼ for a suffix.

If the input word is unknown to Gertwol, we try to segment it with the help of words from our corpus. In addition we use their frequencies in our corpus in order to determine the most likely split. This is done by splitting the compound in every possible way into two parts, so that there are at least three characters left on the left and the right side. The truncated word is then concatenated with each possible right part and the most frequent word, according to our corpus, is taken as solution. In the example *Kuh- und Yakherden* ([herds of] cows and herds of yak) we split the word *Yakherden* and generate the variants *Kuh#herden, Kuh#erden, Kuh#rden* and *Kuh#den* with word parts. Since *Kuh#herden* occurs 11 times in our corpus and all the alternatives do not occur at all, we select this compound for reconstruction and adopt its lemma *Kuh#herde*.

If, in another case, Gertwol provides the information that the input word has more than one strong boundary (#), like for example *Schnee#schuh#touren* in *Ski- und Schneeschuhtouren* (ski [touring] and snowshoe touring), we generate all possible alternatives for the reconstructed lemma by concatenating the truncated word with each of the possible missing parts; *Ski#schuhtour* and *Ski#tour*. Again the corpus frequencies 0 vs. 440 enable us to select the correct lemma *Ski#tour*. We recently realized that this procedure could also help to determine the internal structure of complex compounds, since it will predict that *(Schnee#schuh)#touren* is a more likely interpretation than *Schnee#(schuh#-touren)*. We have not yet explored this idea any further.

5 Manual Error Analysis

During development and testing we encountered typical difficulties with elliptical compounds. In this section we present some of the errors along with ideas on how to solve these problems.

Decision problem: 12.2% of the found cases cannot be decided, which is a drawback of our approach. They may occur in different parts of our lemma reconstructor; if the word has to be split without Gertwol's word segmentations i.e. the word boundaries have to be found automatically, if we have to decide between several solution patterns, and in the decision which part of the compound word has to be attached to the truncated word.

Problematic cases are especially proper names, incl. different spelling variants (e.g. *Monte-Rosa* vs. *Monterosa*), but also old spellings like *Thal* instead of *Tal* (valley) or *Thee* instead of *Tee* (tea), as these are unknown to Gertwol. An approach which could lead to an improvement is the use of a language model to compute the most probable word segmentation. If there exist French translations of the texts, another approach could use the French version of the compound in order to help to solve the German version; e.g. the French version *refuge Vallot, rempli de neige et de salet* could give a useful hint that *schnee- und dreckgefüllte Vallot-Hütte* (Vallot Hut, filled with snow and mud) should result in *schnee#gefüllt* and not in *schnee-#Hütte*.

False Positives **in the search:** Some cases which should not be found, unfortunately have been found. If the part of speech tag accidentally matches a pattern and additionally a wrong "solution" is found by mistake, it may lead to incorrect changes of the lemmas. E.g. *... sieht- und merkbar ...* (... sees and noticeable ...) lead to the "correction" *sieht#bar*. The cause of such problems usually is an incorrectly annotated *TRUNC* tag, hence a tagger error.

***False Negatives* in the search:** There are still cases which our program does not find. On the one hand, due to the lack of coverage in the patterns, on the other hand because of incorrect PoS tags in the corpus, like *bzw.* as *KOUS* instead of *KON* in *Wald- bzw. die Baumgrenze* (forest [line] resp. the tree line), *als* in *Sport- und als Naturschutzverein* (sports [club] and as nature conservation club) as comparative particle *KOKOM* instead of preposition *APPR*, or *haltlose* in *rat- und haltlose Nihilisten* (help[less] and anchorless nihilists) as verb instead of adjective *ADJA*. Furthermore *False Negatives* also occur because of tokenising or other errors, like in *Berg- u.a. Sportarten* (mountain [sport] and other sport) where *u.a.* has been tokenised as one word and annotated with the PoS tag *ADV*. In order to improve the results, the function could be extended by some more patterns.

Very interesting are patterns in which clauses are squeezed in between the truncated word and the elliptical compound. Sometimes the clause is enclosed in parentheses like *Beich- oder* ***(laut der Karte)*** *Birch-grat* (Beich [ridge] or (according to the map) Birch ridge). However, occasionally there is no punctuation to indicate that there is an additional clause in between, e.g. *Knospen- oder* ***für einjährige Gewächse wirklich eine neue*** *Pflanzenwelt* (bud [world] or for one-year-old plants really a new plant world).

Hyphen: As compounds often contain hyphens, like e.g. *Monte-Rosa-Gruppe* (Monte Rosa Massif), our lemma reconstructor handles those cases separately. This means the full compound is divided at the last hyphen, and the last part is attached to the truncated word, which usually works: *Trango- und Biale-Gruppe* (Trango [Massif] and Biale Massif) becomes *Trango#-Gruppe*. However, it leads to mistakes if a word (possibly due to OCR errors or similar) contains hyphens, which should not be there. This is what happened with: *Kohlenhydrat- und Kalo-rienzufuhr* (carbohydrate [supply] and calory supply), which resulted in *Kohlenhydrat#-rienzufuhr*.

Morphology: Sometimes we incorrectly select inflected forms as lemma if the word has been chosen as the lemma of the elliptical compound. Normally, the lemma provided by Gertwol is stored as the lemma, so that it contains word segmentation symbols. If Gertwol does not know the word but there is a lemma specified in our corpus file, the latter is saved. If there is no lemma in the corpus file either, the actual word is stored as lemma, which sometimes results in inflected forms rather than base forms.

Wrong decisions: Errors that occur rarely and usually due to morphology issues are bad decisions by word frequencies. This is the case if, for instance, *Kristall#en* (crystal (dative plural)) with the frequency of 39 instead of *Kristall#fund* (crystal discovery) with the frequency of 3 results as a solution. The same with *ein#en* (one (accusative, masculine)) (17945) instead of *ein#jährig* (one-year-old) (27) or *reg#en* (to move) (129) instead of *reg#los* (motionless) (29). The last two cases were split at the tilde, which is allowed because they are adjectives. Additionally, the lemma choice unfortunately had to be done based on word frequency.

6 Evaluation

We used the 53 German volumes (1957 - 2009) of the Text+Berg corpus' *Release_145_v02*. In total, 11'292 patterns were found. In 1379 cases, i.e. 12.2% of the found patterns, decision problems occurred. As table 1 shows, the simplest pattern is at the same time the most frequent with 7627 occurrences, which are 67.5% of the total cases found. However, this is not surprising concerning the frequency of noun compounds in German. The TIGER corpus also confirms this number with 996 cases of the total of 1362, which is 73.1%.

Freq.	**Pattern**
7627	(9) **TRUNC + KON + NN**
935	(11) WORD + KON + -WORD
522	(6) TRUNC + KON + ADJA/CARD/ADV + NN/NE
488	(2) TRUNC + $, + TRUNC + KON + NN/NE/ADJA
436	(5) TRUNC + KON + ADJ/ADV/VVFIN/VVIZU/VVPP/CARD/APPR + !NN/NE
306	(7) TRUNC + KON + ART/APPR/APPRART + NN/NE
96	(10) TRUNC + APPRART + NN/NE/ADJA
69	(1) TRUNC + $, + TRUNC + $, + TRUNC + KON + NN/NE/ADJA
64	(3) TRUNC + KON + -WORD
33	(8) TRUNC + KON + ART/CARD/ADV/APPR/APPRART + ADJA + NN/NE
29	(4) TRUNC + KON + APPR + ART + NN/NE

Table 1. Frequencies of the found patterns

To determine the accuracy of our lemma reconstructor, we made two different evaluations with small random records of the Swiss Alpine Club yearbooks from 1957 to 2009. On the one hand we evaluated the pattern search (section 4.1) and on the other hand we examined whether the found cases were handled properly (section 4.2). That is, we distinguished between the evaluations for the search and for the corrections.

For the evaluation of the **search** we applied the usual definitions of *precision* ($\frac{TP}{TP+FP}$), *recall* ($\frac{TP}{TP+FN}$) and *f-measure* ($2*\frac{P*R}{P+R}$). For the results of the **correction** we ignored *true negatives* and *false positives* because they are not relevant. In that case, only the ratio of $\frac{corrected\ properly}{correctly\ identified\ as\ TRUNC}$ (*precision*) respectively $\frac{corrected\ properly}{all\ correct\ TRUNC}$ (*recall*) are important. This means that cases which should not be found are ignored, as they are evaluated in the evaluation of the search.

Evaluation of the Search This evaluation was carried out with 106 random sentences, two per book. Additionally we made sure that *TRUNC* tokens occur in each of these sentences, hence the 106 sentences contain 127 *TRUNC* tokens. 65 of which were correctly found by our lemma reconstructor, 56 were correctly not found. Five were not found but should have been found, one was incorrectly found. 49 of the 65 *True Positives* were corrected properly, six changed wrongly and ten were not changed due to decision problems. Furthermore, 50 lemmas were analysed by Gertwol, 15 were segmented with the help of word frequencies.

This means the search evaluation resulted in 98.5% precision, 92.9% recall which corresponds to an f-measure of 95.6%.

Evaluation of the Corrections For this experiment, we took 100 random sentences each of which containing correctly annotated *TRUNC* tokens. In these 100 sentences, 106 *TRUNC* tokens occurred. 100 of them were found by our lemma reconstructor, 89 of which were corrected properly, six changed wrongly and five could not be changed due to decision problems. Six *TRUNC* patterns were not found. Furthermore, 79 lemmas were analysed by Gertwol, 21 were segmented with the help of frequencies. Thus our correction evaluation showed a precision of 89%, a recall of 84%, and an f-measure of 86.4%.

7 Conclusion

We have presented an approach to reconstruct lemmas for truncated words. Our lemma reconstructor analyses the Text+Berg corpus and corrects the lemmas of elliptical coordinated compounds. It looks for specific patterns at the PoS tag level, analyses and splits the compounds and generates solutions depending on the pattern. The splitting of the compounds is supported by a list of words containing word segmentation symbols created by the Gertwol system. The decision between multiple solutions is based on word frequencies which we derived from the corpus itself.

The evaluation of the search reached an *f-measure* of 95.6%, the one of the correction an *f-measure* of 86.4%. Concerning the frequencies of the found patterns, it is remarkable that the simplest construction, namely *TRUNC + KON + NN*, accounts for 67.5% of all matches. The amount of cases with decision problems is 12.2% (1379 of 11'929 found patterns). This shows that the reduction of these cases offers a large potential for improvement, which is planned for future work. This could be achieved, for example, by an additional corpus or by using a language model to compute the most probable word segment which completes the truncated compound. Furthermore, an extension of the search patterns should be taken into consideration to improve the coverage.

The program is open-source, and we are happy to share it with interested parties.

Bibliography

Brants, S., Dipper, S., Hansen, S., Lezius, W., and Smith, G. (2002). The TIGER treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol.

Bredel, U. (2008). *Die Interpunktion des Deutschen.* Niemeyer Verlag.

Eisenberg, P. (2004). *Grundriß der deutschen Grammatik. Der Satz.* Metzler Verlag, Stuttgart, 2. überarbeitete und aktualisierte edition.

Gustafson-Capková, S. and Hartmann, B. (2006). Manual of the Stockholm Umeå corpus version 2.0. description of the content of the SUC 2.0 distribution, including the unfinished documentation by Gunnel Källgren. Technical report, Stockholm University.

Schmid, H. (1994). Probabilistic part of speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

Srinivasan, V. (1993). Punctuation and parsing of real-world texts. In Sikkel, K. and Nijholt, A., editors, *Natural Language Parsing. Methods and Formalisms. ACL/SIGPARSE WORKSHOP. Proceedings of the Sixth Twente Workshop on Language Technology*, pages 163–167, Enschede.

Thielen, C., Schiller, A., Teufel, S., and Stöckert, C. (1999). Guidelines für das Tagging Deutscher Textkorpora mit STTS. Technical report.