University of Zurich[UZH]

Deutsches Seminar

Pluri-Regional German Grammar: A Corpus based Approach

Grammar and Corpora: 4th International Conference, Prague

lic. phil. Simone Ueberwasser

1 The project "Variantengrammatik" ('Variational Grammar')

- A tri-national project (CH, AT, DE), lead by Prof. Dr. Christa Dürscheid, Prof. Dr. Stephan Elspaß, Prof. Dr. Arne Ziegler.
- Six doctoral students.
- Aim: A dictionary that describes regional variation in grammar.
- Restriction: Only regional standards are considered, while dialects are ignored.
- Focus: CH, AT, DE, as well as German speaking parts of Belgium, Luxembourg, Liechtenstein, South Tyrol.
- Supported by SNF, FWF and DFG.
- 2011 – 2014.
- Cf. Dürscheid/Elspaß/Ziegler (2011) and www.variantengrammatik.net
- Stages
  1. Building the corpus
  2. Analyses of the Corpus
     a) Corpus based
     b) Corpus driven
- Building the dictionary

2 The corpus

- Local and regional parts of online newspapers of regional importance (without articles by news agencies, without advertisements).
- 57 newspapers, each with about 5 mio words (≈285 words).
- Aim: Ready by the end of 2012.
- Annotated with
  1. TreeTagger: PoS; lemma.
  2. RFTagger: PoS; grammatical features, such as case, grammatical number, tense etc.
  3. Morphisto: morphological analyses.
  4. Structural units (author, source, date, url, superheading, headline).
  5. Syntactical units (sentence, clause etc.).

The corpus used for the investigation presented here is a subset of the final corpus:

- Version "Work in Progress" (June 2012)
- Total number of corpus texts: 692'547
- Total words in all corpus texts: 250'961'425
- Word types in the corpus: 3'508'115
- Type/token ratio: 0.01 types per token

3 Case studies

3.1 Plural Forms of *Final(e)*

The situation (following the "Variantenwörterbuch"):

- Switzerland:
  - Nom. Sg.: *der Final*
  - Gen. Sg.: *des Finals*
  - Nom. Pl.: *die Finals*
- Germany and Austria:
  - Nom. Sg.: *das Finale*
  - Gen. Sg.: *des Finales* or *des Finals*
  - Nom. Pl.: *die Finale*, *die Finals* or *die Finali*

Query: [lemma = "Finale?" & rfpos = "N.*Sg.*"]

Finale	15'536	98.43%
Finales (Gen.)	248	1.57%
Final	0	0%
Finals (Gen.)	0	0%

Query: [lemma = "Final" & rfpos = "ADJ.*"]

- Searching for *Final* as an adjective finds 1'102 results.
- *Final* (with a lower case <f> unless at the beginning of a sentence) can be an adjective in the sense "the final answer is …".
- TreeTagger recognizes the lemma *Final*, but the RFTagger does not.
- A manual check on 50 randomly selected samples proves 100% of them to be nouns.

Query: [lemma = "Final"]

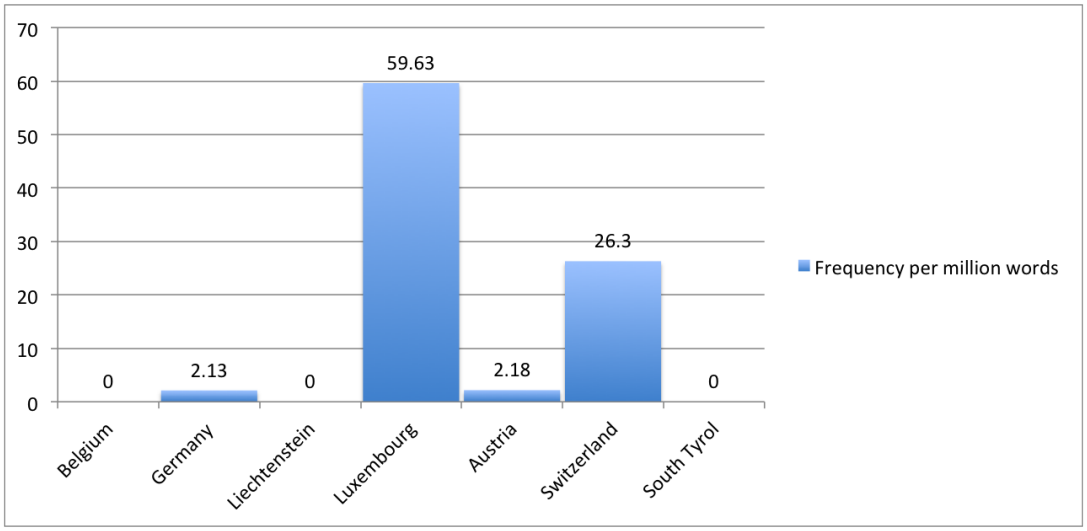

Figure 1: Regional distribution of *Final*.

- The form *Finali* is rare, but defining a search term is no problem.
- The form *Finals* cannot be distinguished from the Gen. Sg. without using articles, which delivers very imprecise results.
- A manual search for [word = "die"][word = "Finale"] discovers individual forms of *die Finale* as plurals. However, finding them systematically is impossible because of the homography with the Sg. form *das Finale*. Only three out of six tokens found are correctly marked as nouns. Only one is recognized as a plural.
- The plural of *Final(e)* is rare all together. The search should be extended to [word = ".*[fF]inale?"] so as to include *semifinal* etc. This decision makes the search even more difficult if the corpus is not tagged for less frequent variants such as *Final*.
- The variant *Final*, which has a low frequency in Germany, is unknown to one of the taggers used.
- Similar problems can be observed for the plural forms *Pärke* (in Germany: *Parks*) and *Spargeln* (in Germany: *Spargel*) and probably many more.

Conclusions

- Taggers were often trained on the most frequent variety only.
- Regional variants are sometimes not recognized by taggers.
- Regular expressions can be used to search for word forms. However:
  - This is not feasible in case of homography of word forms (e.g. Gen. Sg. *Finals* and Nom. Pl. *Finals*).
  - If the PoS tagging is wrong (e.g. Adjective for nouns), grammatical annotations are lost (case, numerus etc.).
  - These wrong taggings have an influence on syntactic searches, too.
  - These wrong taggings might disturb corpus-driven research.

Excursion:* Final *in the IDS corpus

- TAGGED-C - Archiv morphosyntakt. annotierter Korpora (CONNEXOR)
  - *Final* is marked correctly as a noun.
- TAGGED-T - Archiv morphosyntakt. annotierter Korpora (TreeTagger)
  - *Final* is marked as an adjective.
- TAGGED-M - Archiv morphosyntakt. annotierter Korpora (MECOLB; ehem. TAGGED)
  - Not enough data to tell, especially no Swiss data.

For research on variation, taggers have to be trained for less frequent variants.

3.2 Case selection after *wegen*

The situation (following "Zweifelsfälle Duden"):

- The preposition[1] *wegen* ('because of') can take a Genitive or a Dative.
- The use of the Genitive is the standard. Use of the Dative is regional and/or substandard in written language.

The questions:

- If the use of the Dative is very frequent in a specific region, can it still be considered to be non-standard there?
- What frequency does it take to support such an assumption?

[1] *wegen* can also be used as a postposition, a variant that certainly has to be investigated but will be neglected here.

[lemma = "wegen"][rfpos = "ADJD.*"]*[rfpos = ".*Dat.*"] within c
[lemma = "wegen"][rfpos = "ADJD.*"]*[rfpos = ".*Gen.*"] within c [2]

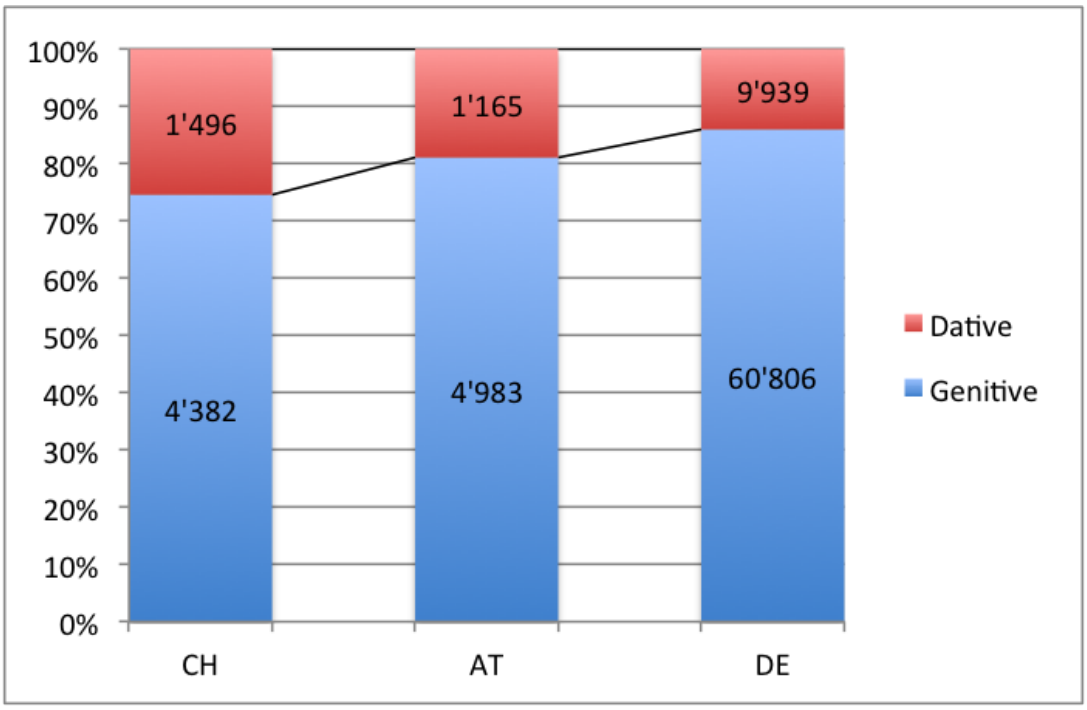

Figure 2: Distribution of *wegen* with Dative and Genitive respectively in Switzerland, Austria and Germany [%]

Use of the Dative is more frequent in Austria and Switzerland. But how reliable are the data used?

Manual counting of 100 occurrences marked as Dative by the RFTagger

Dative	53%
Syncretism [3]	45%
Genitive	2%

Manual counting of 100 occurrences marked as Genitive by the RFTagger

Syncretism	53%
Genitive	47%

Around 50% of the occurrences of the objects of* wegen *are not annotated with the case, in which they appear.

[2] An additional 571 occurrences are marked as unknown cases by the RFTagger.

Possible reasons for the low reliability

- Short style of newspapers: in *wegen Bauarbeiten* the case is not visible. If an article had been used (*wegen der*$_{Gen}$ *Bauarbeiten* / *wegen den*$_{Dat}$ *Bauarbeiten*), the situation would be clear and the tagger would probably have performed better.
- The tagger not only uses the form of the noun phrase but also the environment to define the case. This also explains the slightly higher precision when detecting Datives, since the preposition *wegen* is seen as preferring the Genitive, so unclear noun phrases are more likely tagged in this case. Ideally, noun phrases would be tagged with both, the form of the case (i.e. the visible morphology) but also the category (i.e. the case as it depends on the preposition, verb etc. that governs the noun phrase). However, taggers do not work on this level.

Possible work-arounds

- Take a smaller sample, count manually and extrapolate.
- Find means to restrict the search to those noun phrases which do show recognizable features of either case, i.e. noun phrases with articles, adjectives, nouns ending in <-s> etc.

  Either approach, however, has its pitfalls.

100 randomly selected* wegen *clauses per country
([lemma = "wegen"][rfpos = "ADJD.*"]*[] within c)

	CH	AT	DE
Dative	14	7	4
Genitive	44	36	44
Syncretism	38	54	47
null[4]	4	3	5
Total	100	100	100

100 randomly selected* wegen *clauses per country, recognizable cases only

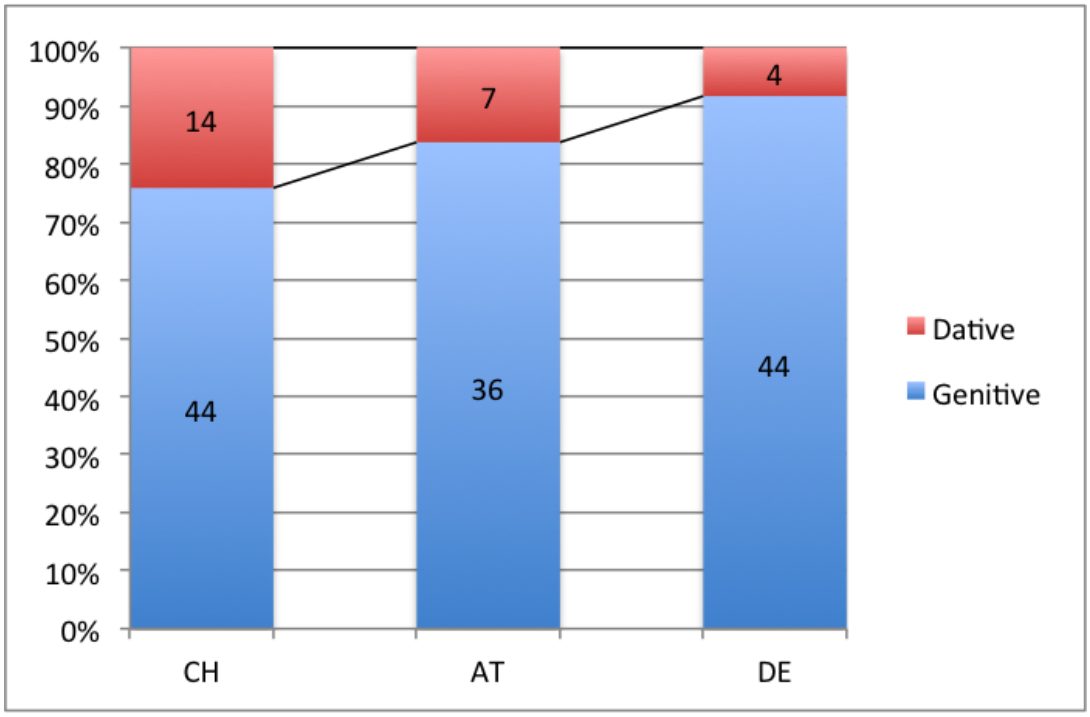

Figure 3: Distribution of *wegen* with Dative and Genitive respectively in Switzerland, Austria and Germany in 100 randomly selected *wegen*-clauses from either country [%].

- The significance of this result is too low to be acceptable (χ^2 = 4.709, df = 2, p = 0.09494097).
- A clear tendency can be seen, but larger samples are needed.
- The distribution in nearly equal to the one received with the incorrect case tagging. However, this statement has no value, since data that are not significant can not be used to support a theory based on wrong data.
- Manual counting of a bigger sample that delivers significant results is absolutely necessary.

4 Conclusions

- Taggers were trained on data that were considered as standard at the time and place of training. Investigating data outside this spectrum poses problems, which so far have been recognized for diachronic corpora but not for diatopic ones.
- Consequently, when investigating data, researchers have to pay special attention to the nature and tagging of the corpus. A corpus that was created and adjusted explicitly for the purpose of the study at hand might be a better solution than a read-made corpus.
- When working with corpora, one has to be aware of this type of problems and look for work-arounds, which can still comply with the needs for accuracy and reliability.
- From the point of view of the person working with the corpus, the option to see the actual tagging is an important desire (cf. the noun *Final* which is tagged as an adjective).

5 Bibliography

Ammon, Ulrich et al. (2004): Variantenwörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz und Deutschland sowie in Liechtenstein, Luxemburg, Ostbelgien und Südtirol. Berlin, New York: de Gruyter.

Dudenverlag (Hrsg.) (2011): Richtiges und gutes Deutsch. Wörterbuch der sprachlichen Zweifelsfälle (7., vollständig überarbeitete und erweiterte Auflage) (=Der Duden in zwölf Bänden 9). Mannheim: Bibliographisches Institut & F. A. Brockhaus.

Dürscheid, Christa et al. (2011): Grammatische Variabilität im Gebrauchsstandard: das Projekt „Variantengrammatik des Standarddeutschen“. In: Konopka, Marek et al. (Hrsg.): Grammatik und Korpora 2009. Dritte Internationale Konferenz. Mannheim, 22.-24.09.2009. (=Korpuslinguistik und interdisziplinäre Persepektiven auf Sprache 1. Hrsg: Keibel, Holger et al.). Tübingen: Narr, 123-140.