Published in: Gurevych, Iryna/Biemann, Chris/Zesch, Torsten (eds.): Language Processing and Knowledge in the Web. 25th International Conference, GSCL 2013, Darmstadt, Germany, September 25-27, 2013. Proceedings. - Berlin/Heidelberg: Springer, 2013. pp. 83-88. (Lecture Notes in Computer Science 8105)

POSTPRINT

Decision Tree-Based Evaluation of Genitive Classification – An Empirical Study on CMC and Text Corpora

Sandra Hansen and Roman Schneider

Institute for German Language (IDS), Mannheim/Germany
{hansen,schneider}@ids-mannheim.de

Abstract. Contemporary studies on the characteristics of natural language benefit enormously from the increasing amount of linguistic corpora. Aside from text and speech corpora, corpora of computer-mediated communication (CMC) position themselves between orality and literacy, and beyond that provide insight into the impact of "new", mainly internet-based media on language behaviour. In this paper, we present an empirical attempt to work with annotated CMC corpora for the explanation of linguistic phenomena. In concrete terms, we implement machine learning algorithms to produce decision trees that reveal rules and tendencies about the use of genitive markers in German.

Keywords: Corpus Linguistics, Computer-Mediated Communication, Machine Learning, Decision Trees, Grammar, Genitive Classification.

1 Introduction

Linguistic studies are increasingly corpus-based, i.e. their statements rely on empirical data, computed on the basis of natural language. Due to the problematic nature of speech corpora, e.g. the difficulty of achieving substantial amounts of authentical spoken samples or the influence of situational conditions on the proband's speech behavior, text corpora currently represent the vast majority of available resources. In this situation, corpora of computer-mediated communication (CMC) open up new possibilites for the examination of language phenomena between the poles of orality and literacy [2] [5]. Internet-based discourse genres such as e-mails, weblogs, or chat and discussion groups offer insight into the use of language in situations that are at least to some extent close to verbal data and face-to-face communication [7].

It is well known that, due to specific production conditions, the syntactical rules of spoken and computer-mediated language differ from the rules that apply to written language. This has substantial impact on the performance of linguistic tools like taggers and parsers. Therefore it seems most desirable to verify the conditions under which automatically annotated CMC corpora can contribute to linguistic research. We are especially interested in the question whether the statistical evaluation of hypotheses based on machine learning algorithms is applicable. For a first estimate, we compare the results of an empirical study conducted on the basis of a large text corpus with the output of the same methods and algorithms adapted to CMC data.

The evaluation of hypotheses predicting the use of genitive markers in German is a field of study that is notoriously complicated and generates cases of doubt, because there is no generally accepted model: Is it better to use "des Films" or "des Filmes" (i.e., to use "-s" or "-es" marker)? Under which conditions is it tolerable to omit the genitive marker (e.g., zero-marker as in "des Internet")? In order to find an empirical answer, manifold intra- and extralinguistic parameters has to be considered: the number of word syllables, types of coda, noun frequency, information about medium, register, and region etc. Therefore, decision trees seems to be a valuable tool to identify, order, and structure the factors that are most prominent for the actual decision.

2 Corpus Resources

For our study, we used the *Dortmund Chat Corpus*[1] that was compiled between 2003 and 2009. It covers logfiles from different chat groups, supplemented with CMC-specific metadata and encoded in an interchangeable XML format, ranging over a variety of subjects and situational contexts. Though the complete corpus contains more than one million word forms, the publicly available release has to content itself with 548,067 word forms within 59,558 chat postings. For our further processing, the original chat texts were annotated morphosyntactically with three competing systems: Connexor Machinese Tagger, TreeTagger, and Xerox Incremental Parser[2]. In the following, we primarily use the Xerox parser because it gives us the broadest range of syntactic and structural annotation, for example case information for nouns. As text-oriented counterpart, we choose the 2011-I release of the *German Reference Corpus DeReKo*[3] with more than 4 billion word forms, which is one of the major resources worldwide for the study of written German. Like for the CMC corpus, morphosyntactic annotations from the three tools mentioned above are added, and the corpus is enriched with a comprehensive set of extra-linguistic metadata. Language samples, annotations, and metadata were integrated into a prototypical RDBMS-driven corpus storage and retrieval framework. This system allows for the flexible analysis of multi-layered corpora with regular expressions and a combined search on all available types of annotation and metadata, using parallelized SQL queries and a MapReduce-like retrieval paradigma. Our study benefits from the fact that within the framework all language samples are stored wordwise, and every wordform is connected to intra- and extra-linguistic metadata according to an efficient logical data model [9].

3 The Genitive Extraction

The primary corpus data served as a basis to extract all relevant genitive forms. As a first step, the genitive candidates were filtered out using a specifically adjusted Perl script. The resulting database consisted of about 454,500 types and 7,334,500 tokens.

---

[1] `http://www.chatkorpus.tu-dortmund.de`

[2] See `http://www.connexor.eu/technology/machinese/index.html`, `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`, and `http://open.xerox.com/Services/XIPParser/`, respectively.

[3] `http://www.ids-mannheim.de/DeReKo/`

Then, by means of the script and in order to weight the findings, several distribution rules were checked. For example, in cases where the word ends with a genitive marker and the lemma does not end with a marker, the genitive candidate gets a so-called score point. In the context of a second distribution rule, we give an additional score point for a typical pre- or post-modified genitive preposition. If the script detects an adjacent genitive article or a genitive article within a certain distance from a premodificated, adjacent adjective, it assumes the presence of a noun with a genitive form, and the token in question gets two more score points. The following example shows a genitive noun (token = "*Anblicks*"; lemma = "*anblick*") with a genitive preposition ("*wegen*") followed by a genitive article ("*des*") and a premodificated adjective ("*schönen*"): "*wegen des schönen Anblicks*".

Overall, we implemented 19 different distribution rules, and counted the total of the assigned score points for every genitive candidate. The higher the score points, the more likely the candidate was considered a genitive noun. All candidates with score points greater than two were taken into account. The script output was measured against a manually annotated gold standard, containing 9,000 nouns extracted out of 1,000 sentences. Precision, recall, and F-scores are about 95%.

Within a following step, the candidates were enriched with metadata (location, medium, domain, year etc.) and morphosyntactic information in order to get additional grammatical evidence (e.g., information about the genus). We isolated loanwords, acronyms, and neologisms using existing word lists. Some distributionally motivated information was extracted with a second Perl script. By comparing our data set with CELEX [1], we were able to include phonetic and prosodic information (e.g., the number of syllables or the character of the last sound/coda) into our calculations.

4 Statistical Analysis

Subsequently, we evaluated the factors influencing the use of genitive markers, and tried to model a decision tree for both corpora based on the token data. We started by encoding 34 language-immanent and extra-linguistic factors influencing the marking of a genitive noun. To get a general idea about the specific factor's influences and side effects, we calculated chi-square-tests and visualized the residuals with an association plot (cf. [3] [4] [6])[4]. The plots visualize the standard deviations of the observed frequencies as a function of the expected frequencies. Each cell is represented by a rectangle, whose height is proportional to the residual of the cell, and having a width proportional to the square root of the expected frequency. Therefore, the area of the rectangle is proportional to the difference between observed and expected frequencies.

As an in-depth presentation of all factors would exceed the limits of this paper, we will concentrate on a rather small selection. Figure 1 represents the influence of the number of syllables to the genitive marker in the DeReKo-based text corpus. The association plot shows the under-representation of the "ns"-, zero-, "s"- and "ses"-markers and the over-representation of the "ens"- and "es"- markers as a function of the number of syllables. If the lexeme consists of multi syllables, the "ens"- and "es"-markers are under-represented and the residuals of the "s"-markers are much higher.

[4] All tests and plots were conducted and produced using the VCD package (Visualizing Categorial Data) of the statistical software "R".

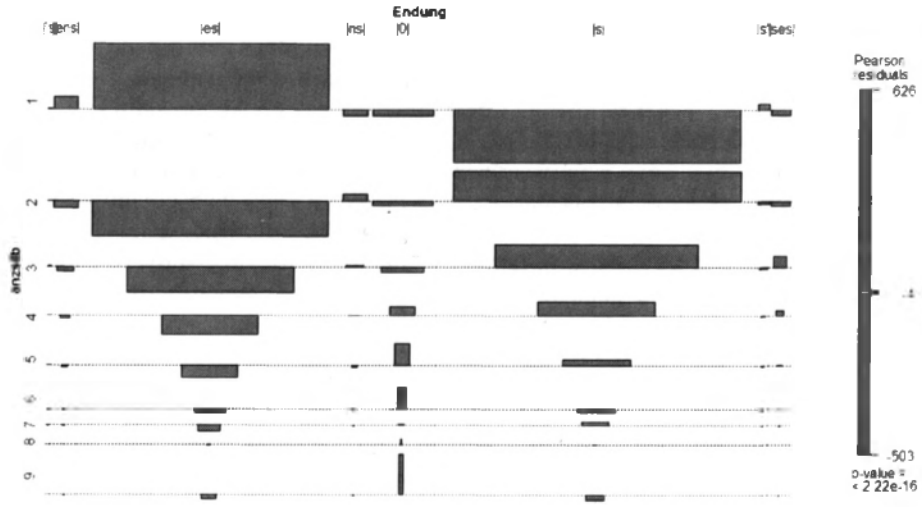

Fig. 1. Association plot for the influence of number of syllables (text corpus)

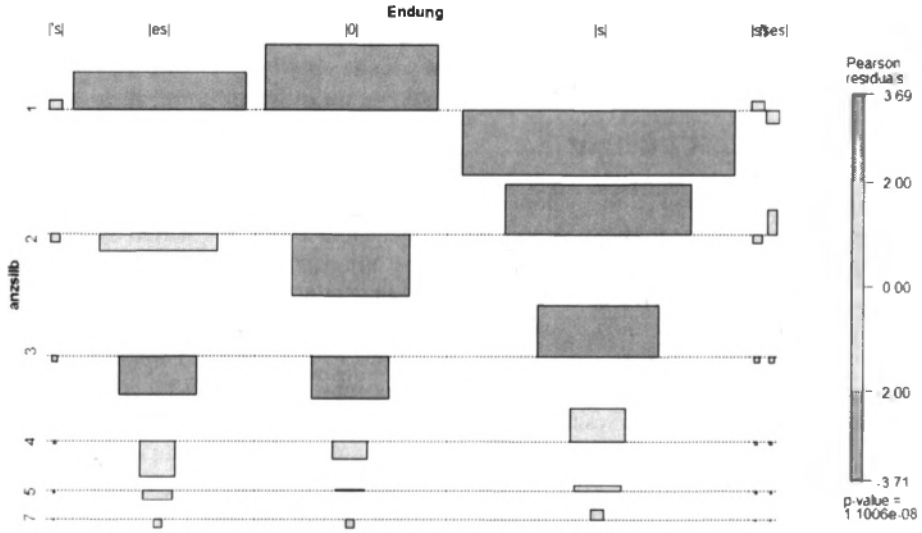

Fig. 2. Association plot for the influence of number of syllables (CMC corpus)

Figure 2 shows the same statistical analysis on the CMC data set. Concerning the "es"- and "s"-markers, the results are similar to those of the text corpus. But interestingly, the residuals of the zero-marker (value "0" in the plots above) show different trends. Within the chat corpus, the zero-marker of words with one syllable is strongly over-represented, whereas the zero-marker in multi syllable words is significantly under-represented. Here, some in-depth linguistic interpretation would be valuable.

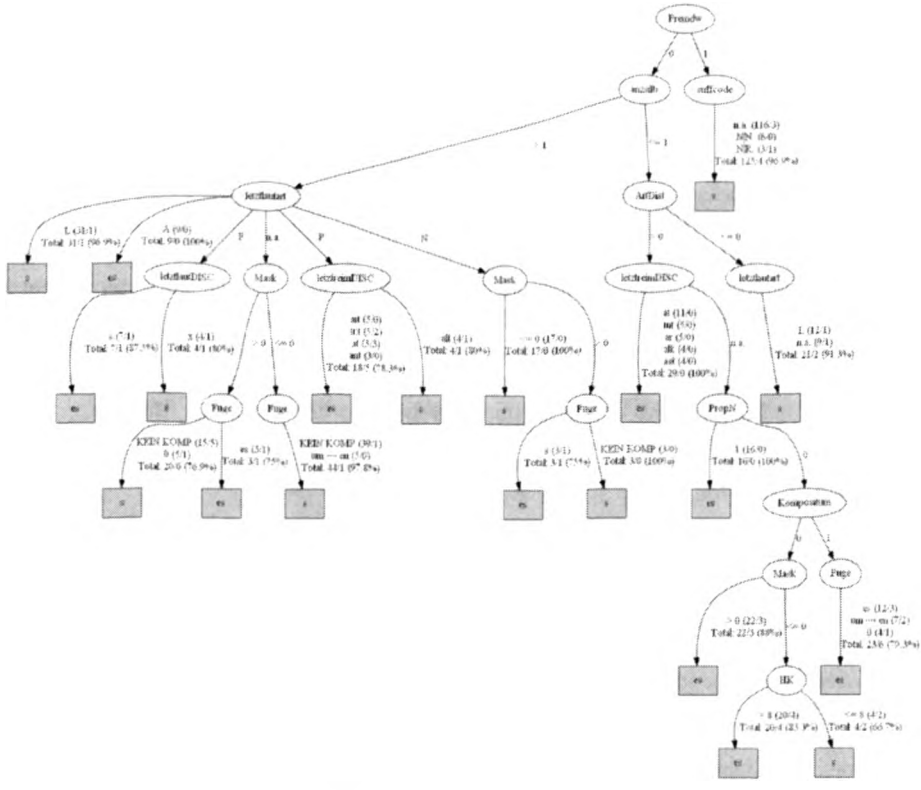

Fig. 3. Decision tree with C 4.5 factors predicting the use of genitive markers

The association plots and chi-square tests were produced and conducted for every single factor in order to test their influence on the distribution of genitive markers.

Afterwards, the main analysis was focused on the calculation and visualization of complex decision trees for each corpus. Decision tree learning uses tree representations as a predictive model that maps observations about an item to conclusions about the item's target value. The leaves of the trees represent class labels, the branches represent conjunctions of features that lead to these class labels.

For the statistical analysis, we took recourse to the C4.5 algorithm [8] using the WEKA software [10]. We calculated decision trees for both the chat corpus and the text corpus in order to compare the various relationships of the factors predicting the use of genitive markers. As a result, 91.35% of the instances in the text corpus were correctly classified. In the CMC corpus, the correctly classified instances are 90.89%.

Figure 3 displays the decision tree for the prediction of genitive markers in the chat corpus. It has to be read as follows: The first significant single factor splitting the data into separate groups is whether the noun is – or is not – a loanword (leaf "Fremdw"). Then, if the lexeme is part of the basic vocabulary (branch "0"), the factor "number of syllables" (leaf "anzsilb") is relevant, and so on. The complete decision tree for the text corpus is far too large and comprehensive to be included in this paper, but it can

be accessed online.[5] It shows some notable differences, as for written texts the frequency of a lexeme seems to be a highly relevant factor just at the top of the tree.

5 Summary and Outlook

We presented a novel empirical approach to work with annotated CMC corpora for the explanation of linguistic phenomena, using the example of German genitive markers. We used machine learning algorithms to produce decision trees showing differences between a CMC corpus and a "traditional" text corpus. They reveal that a lot of the most influential factors predicting genitive marking are the same, but also that the sequences and the interaction of the factors are different. We will further investigate the involved mechanisms, verify the reliability of automated annotations for CMC data, and explore the linguistic interpretations for the statistical findings.

References

1. Baayen, R.H., Piepenbrock, R., Gulikers, L.: The CELEX Lexical Database (CD- ROM), Philadelphia (1995)
2. Beißwenger, M., Storrer, A.: Corpora of Computer-Mediated Communication. In: Lüdeling, A., Kytö, M. (eds.) Corpus Linguistics, vol. 1, pp. 292–308. de Gruyter, Berlin (2008)
3. Cohen, A.: On the graphical display of the significant components in a two-way contingency table. In: Communications in Statistics - Theory and Methods, vol. A9, pp. 1025–1041 (1980)
4. Friendly, M.: Graphical methods for categorical data. In: SAS User Group Int. Conference Proc., vol. 17, pp. 190–200 (1992), http://www.math.yorku.ca/SCS/sugi/sugi17-paper.html
5. Herring, S.: Computer-Mediated Conversation. Language@Internet 7/8 (2010/2011), http://www.languageatinternet.org
6. Meyer, D., Zeileis, A., Hornik, K.: The strucplot framework: Visualizing multi-way contingency tables with vcd. Report 22, Department of Statistics and Mathematics, Wirtschaftsuniversität Wien, Research Report Series (2005)
7. Ogura, K., Nishimoto, K.: Is a Face-to-Face Conversation Model Applicable to Chat Conversations? In: Proc. PRICAI Workshop Language Sense on Computer, pp. 26–31 (2004)
8. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Francisco (1993)
9. Schneider, R.: Evaluating DBMS-Based Access Strategies to Very Large Multi-Layer Annotated Corpora. In: Proceedings of the LREC-2012 Workshop on Challenges in the Management of Large Corpora, Istanbul (2012)
10. Witten, I., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, San Francisco (2005)

[5] The tree can be accessed here:
http://hypermedia.ids-mannheim.de/treeText.c095.m2000.pdf