GenitivDB – a Corpus-Generated Database for German Genitive Classification

Roman Schneider
Institute for German Language (IDS)
R5 6-13, D-68161 Mannheim
E-mail: schneider@ids-mannheim.de

Abstract

We present a novel NLP resource for the explanation of linguistic phenomena, built and evaluated exploring very large annotated language corpora. For the compilation, we use the *German Reference Corpus (DeReKo)* with more than 5 billion word forms, which is the largest linguistic resource worldwide for the study of contemporary written German. The result is a comprehensive database of German genitive formations, enriched with a broad range of intra- und extralinguistic metadata. It can be used for the notoriously controversial classification and prediction of genitive endings (short endings, long endings, zero-marker). We also evaluate the main factors influencing the use of specific endings. To get a general idea about a factor's influences and its side effects, we calculate chi-square-tests and visualize the residuals with an association plot. The results are evaluated against a gold standard by implementing tree-based machine learning algorithms. For the statistical analysis, we applied the supervised LMT Logistic Model Trees algorithm, using the WEKA software. We intend to use this gold standard to evaluate *GenitivDB*, as well as to explore methodologies for a predictive genitive model.

Keywords: NLP, Metadata, Grammar

1. Motivation

Genitive is the grammatical case used to indicate a type of relationship – most prominently possession/ownership – between things. In German as well as in English, genitive nouns within a genitive phrase are identifiable by certain endings. But whereas English features just two variants (singular nouns: add apostrophe S (*Peter's*), plural nouns ending in *-s*: only add the apostrophe (*sisters'*)), the German language morphologically marks genitive constructions with up to eight distinct variants. Examples for German genitive nouns with different types of markers are displayed in table 1.

Lemma	Genitive case	Type of marker
Mutter	*Mutters*	*-s*
Kind	*Kindes*	*-es*
Gedanke	*Gedankens*	*-ns*
Bedürfnis	*Bedürfnisses*	*-ses*
Herz	*Herzens*	*-ens*
Peter	*Peter's*	*-'s*
Felix	*Felix'*	*-'*
Internet	*Internet*	*(zero-marker)*

Table 1: Types of genitive markers in German

Research on the genitive case in German has a long tradition. Nevertheless the evaluation of hypotheses predicting the use of certain genitive variants is notoriously complicated and generates cases of doubt, because there is no generally accepted model. Examples for common questions are:

- Is it better to use "*des Films*" or "*des Filmes*" (i.e., to use the "*-s*" or "*-es*" marker)?
- Is the marking of genitive forms of proper names with apostrophe S instead of using the more traditional *-s* ending (i.e., "*Peter's*" instead of "*Peters*") really good grammatical style?
- Under which conditions is it tolerable to omit the genitive marker (e.g., zero-marker as in "*des Internet*")?

In order to find answers, up to 30 intra- and extralinguistic factors have been considered in the past: position of the genitive attribute, article ending in *-s*, morphological complexity, number of syllables, types of coda, lexical integration, genus, geograhic or proper name, noun frequency, information about medium, register, and region etc. (Fehringer, 2011; Szczepaniak, 2010). On these grounds, it seems difficult to define a consistent model and to identify weighting criteria – statistically spoken: the effect size – for certain factors.

Over the decades, several hypotheses were proposed. Just to name a few: Appel (1941) postulates that the omission of genitive markers affects primarily special/technical vocabularies; Pfeffer & Morrison (1984) describe – among other aspects – the influence of the final syllable; Fehringer (2011) and Szczepaniak (2010) assume that markers are determined by the number of noun syllables, the frequency of the genitive noun, etc. Standard references like Dudenredaktion (2007) and Dudenredaktion (2009) try to present the classification problem in its entirety.

Contemporary studies on the characteristics of natural language benefit enormously from the increasing amount of linguistic corpora. Linguistic findings are increasingly corpus-based, i.e. their statements rely on empirical data, computed on the basis of natural language. However, resources for the multifactorial examination of genitive formation are scarce. We thus present a novel corpus-based data collection and a statistical approach to identify, order, and structure the factors that are most prominent for genitive variation in German. This is a first step towards a comprehensive description of genitive variation based on actual language use.

2. Corpus Resources

For the compilation of the genitive database *GenitivDB*, we use the *German Reference Corpus DeReKo*[1] with more than 5 billion word forms, which is the largest linguistic resource worldwide for the study of written German. The original texts are annotated morphosyntactically with three competing systems: Connexor Machinese Tagger, TreeTagger, and Xerox Incremental Parser[2]. In the following, we primarily make use of the Xerox and TreeTagger annotations because they give us the broadest range of syntactic and structural annotation – case information for nouns, phrase boundaries etc. – as well as reliable lemmatizations. Besides, the corpus is semi-automatically enriched with a comprehensive set of metadata (text type, year of publication, regional background, topic, medium, etc.). Language samples, annotations, and metadata are fully integrated into a RDBMS-driven corpus storage and retrieval framework. This corpus database, *KOGRA-DB* (Schneider, 2012), 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 separation of genitive variants benefits from the fact that all language samples are stored wordwise, and that every wordform is connected to intra- and extra-linguistic metadata according to an efficient logical data model.

3. Building the Genitive Database

The corpus data serve as a basis to extract all relevant genitive forms. Potential candidates are filtered out using regular expression queries on the primary texts and metadata. After several refinements, the resulting collection comprises 650,726 types and 9,541,753 tokens. The most prominent ending is *-s*, followed by *-es* (see figure 1 for relative frequencies). In order to weight the findings, several distribution rules are checked automatically, e.g.:

- If the wordform ends with a genitive marker (*-ens, -es, -ns, -s, -ses*) and its lemma does not end with a marker, the genitive candidate gets a so-called score point.
- We give an additional score point if the candidate is pre- or postmodificated by a genitive preposition.
- If our script detects an adjacent genitive article in front of the noun, the candidate gets two more score points.
- If we find a genitive article within a certain distance in front of the noun and an inflected premodifying adjective ending in *-en* , a proper name form in *-er*, or an ordinal number immediately in front of the noun, it also 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 premodifying adjective ("*schönen*"): "*wegen des schönen Anblicks*".

Overall, we make use of 19 different distribution rules, and count the total of the assigned score points for every genitive candidate. The higher the score points, the more likely the candidate can be considered a genitive noun. A final manual inspection suggests that all candidates with two or more score points are "real" genitive variants, whereas the others become weak candidates.

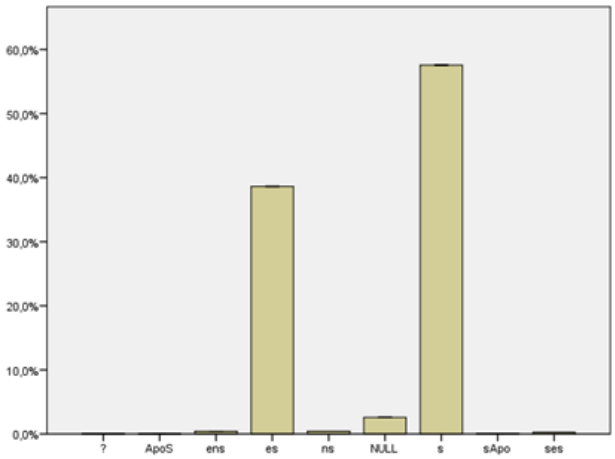

Figure 1: Relative freqency of German genitive markers

All findings are enriched with extra-linguistic metadata and morphosyntactic information from *KOGRA-DB* in order to get additional grammatical evidence. We isolate loanwords, acronyms, and neologisms using existing word lists from in-house projects.[3] Some distributionally motivated information is added with a specific Perl script. By matching our dataset against *CELEX* (Baayen et al., 1995), we are also able to include phonetic and prosodic information (e.g., the number of syllables or the character of the last sound/coda) into our calculations.

Subsequently, we evaluate the main factors influencing the use of genitive markers (see also Hansen & Schneider, 2013). To get a general idea about a specific factor's influences and side effects, we calculate chi-square-tests and visualize the residuals with an association plot (cf. Cohen, 1980; Meyer et al., 2005), using the VCD (Visualizing Categorial Data) package of the statistical software "R". The plots show 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.

[1] *http://www.ids-mannheim.de/DeReKo*

[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] These lists were compiled by the projects *OWID* (*http://www.owid.de*) and *Deutsches Fremdwörterbuch* (*DFWB; http://www.ids-mannheim.de/lexik/fremdwort.html*).

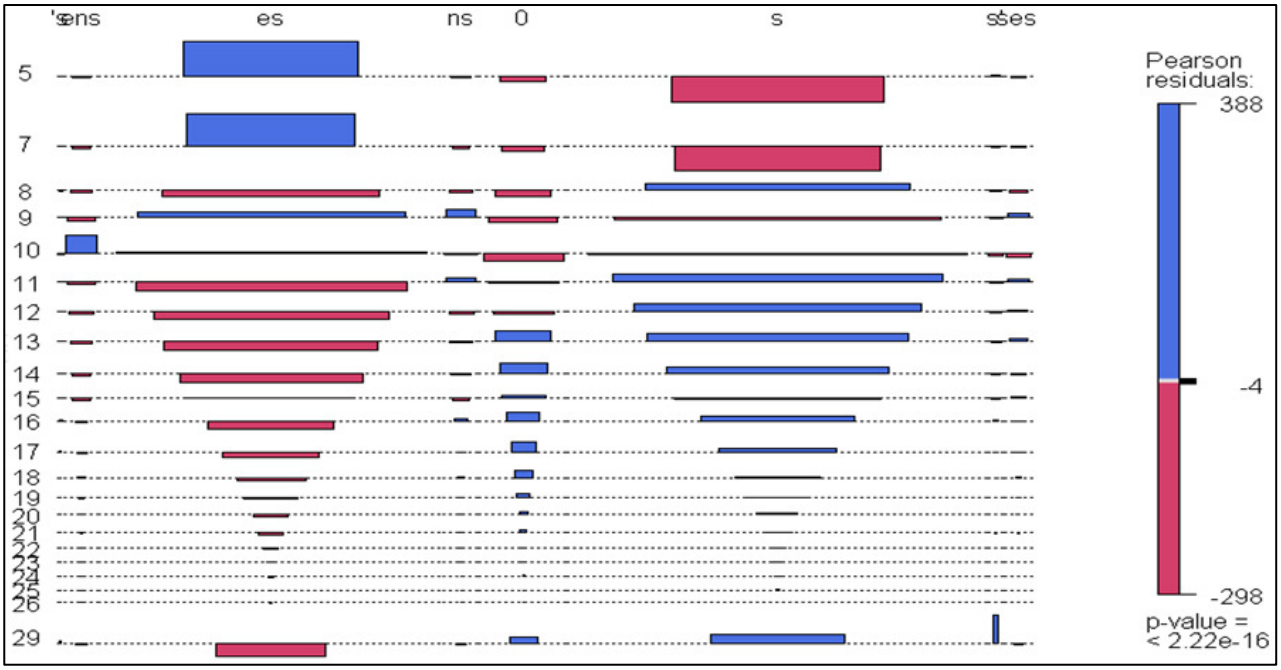

Figure 2: Influence of word frequency classes (5-29) on genitive formation

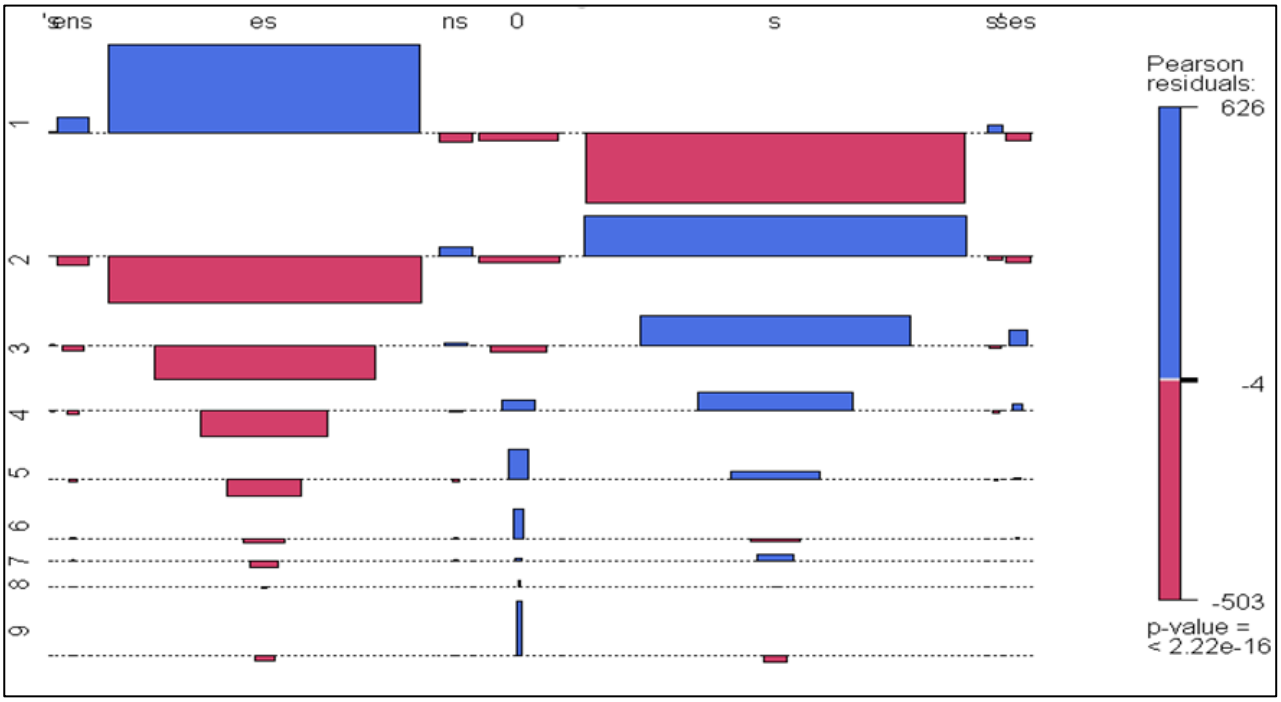

Figure 3: Influence of the number of syllables (1-9) on genitive formation

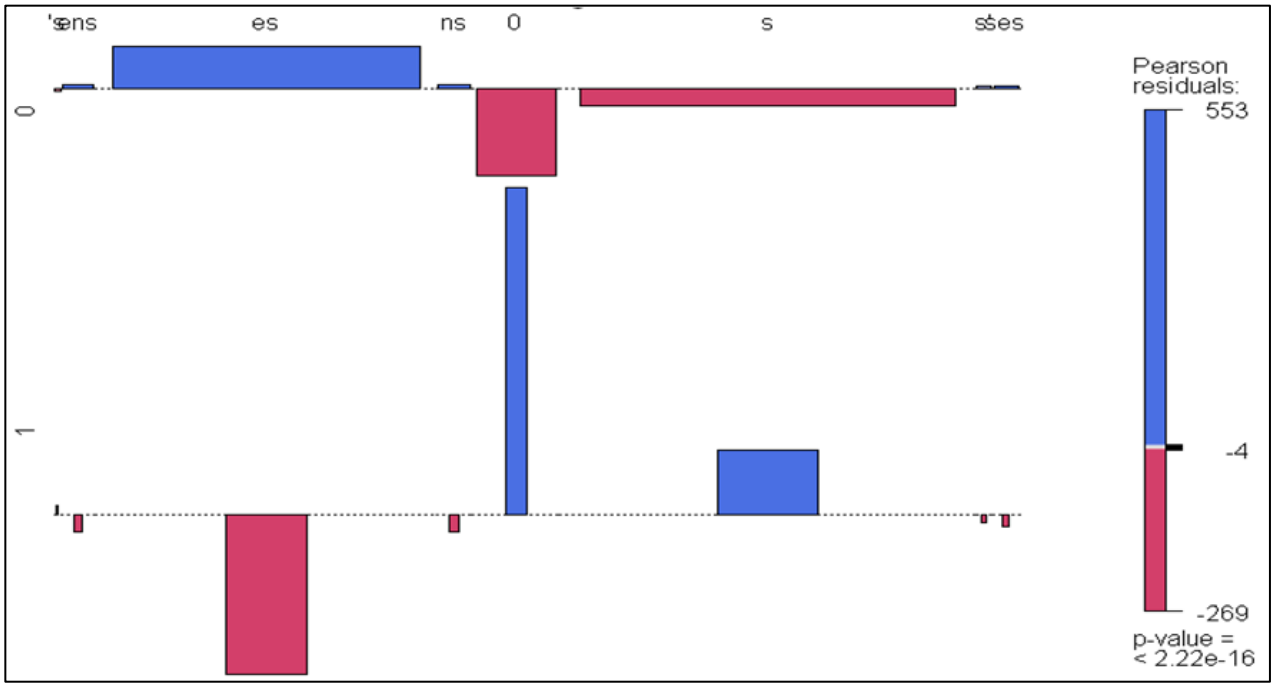

Figure 4: Influence of loanword status (0/1) on genitive formation

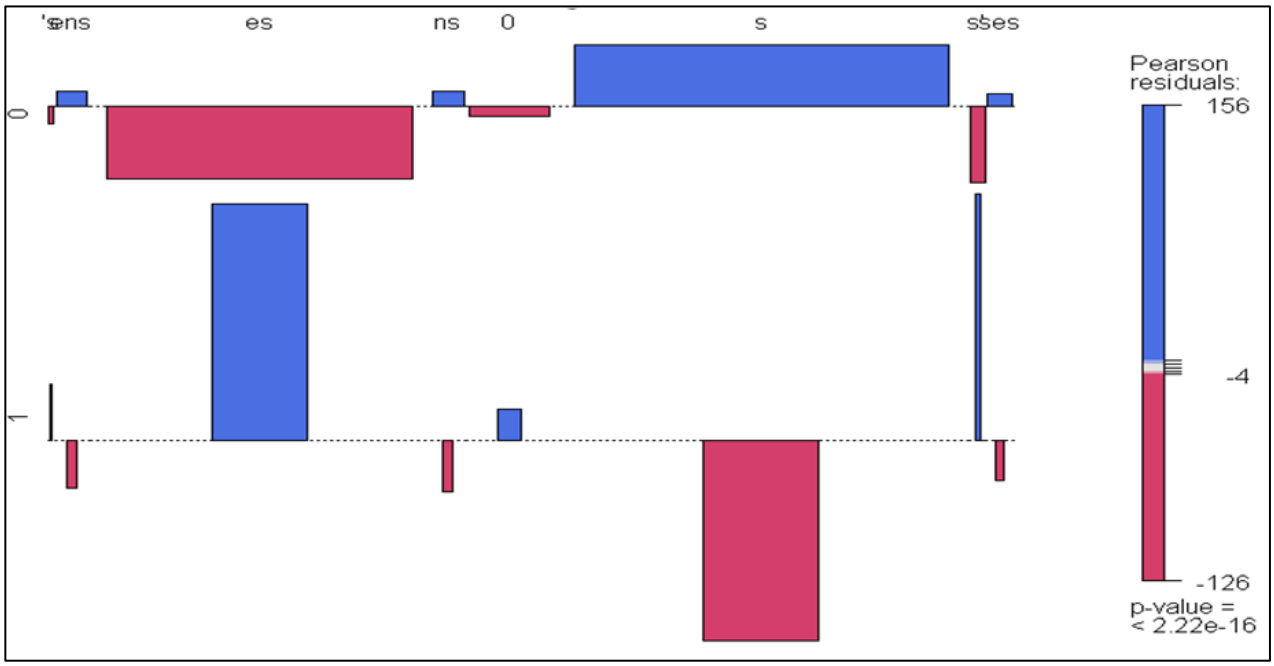

Figure 5: Influence of proper noun status (0/1) on genitive formation

As an in-depth presentation of all factors would exceed the limits of this paper, we concentrate on a rather small selection. Figure 2 represents the influence of the frequency class[4] (vertical y-axis) on genitive formation. The association plot shows the under-representation of *s*- and zero-markers and the over-representation of *es*-markers for nouns with higher frequencies (frequency class up to 7 or 10, respectively). So the representation of genitive markers as a function of the word frequency reveals a trend for preferring the *-es* variants for words that are often used in contemporary German – an empirical validation for a widespread assumption in linguistic research. Figure 3 displays the influence of the number of noun syllables. It shows that the genitive ending *-es* is over-represented for nouns with one syllable, whereas multisyllabic words tend to use the shorter *-s* variant. Figures 4 and 5 indicate an inverse influence of loanword status and proper noun status on the use of *-es* and *-s* markers, as well as a significant preference of loanwords for zero-markers.

In addition to this quite straightforward first examination, several other factors and their combinations are worth further investigation. The association plots and chi-square tests produced and conducted for every single factor in our dataset constitute a valuable basis for the description of their influence on the distribution of genitive markers.[5] Some of the most significant parameters that we also use for the multifactorial evaluation of our gold standard within the next section are:

- Proper noun (*yes/no*)
- Adjacent adjective ending in *-en* in front of the noun (*yes/no*)
- Adjacent noun (*yes/no*)
- Neologism (*yes/no*)
- Loanword (*yes/no*)
- Compound word (*yes/no*)
- Genus (separate probability values for *fem, masc, neut*)
- Frequency class (*1-29*)
- Parser output (TreeTagger) on genitive probabilty (decimal number)
- Domain (*fiction, cultural/entertainment, nature, technology, politics/society*)
- Medium (*press, books, internet, spoken*)
- Location (we use eight greater regions from the German-speaking area)

We store the complete dataset (9,541,753 genitive candidates with sentence context and more than 80 qualified attributes for each wordform) within a relational database management system. This resource (*GenitivDB*) is well-documented and can be queried online. Query parameters of the frontend are lemma, wordform, genitive marker, and genitive probability based on the assigned score points for every genitive candidate (see figure 8). We also provide a complete public download for further scientific exploration.[6]

4. The Gold Standard

In order to answer the demand for a gold standard set for genitive classification, we manually inspect over a set of 1,000 randomly chosen sentences from *KOGRA-DB*.[7] All singular nouns (>9,000) are checked for genitive markers, genus, and inflection class (*weak/strong*) by native speakers with linguistic background. This results in about 300 verified genitive nouns, to which we add the metadata described above. We intend to use this gold standard to evaluate *GenitivDB*, as well as to explore methodologies for a predictive genitive model.

For the statistical analysis, we applied the supervised LMT Logistic Model Trees algorithm (Landwehr et al., 2005) using the free WEKA (Waikato Environment for Knowledge Analysis) machine learning workbench (Witten & Frank, 2005). Figure 6 documents the header and data sections of our ARFF (Attribute-Relation File Format) import file. The first line declares the relation, followed by 16 lines introducing the complete attribute list with type information. The (truncated) data section contains some generic instances with the genitive marker at the end of each row.

```
@relation genitive
@attribute fem numeric
@attribute masc numeric
@attribute neut numeric
@attribute morphgen numeric
@attribute nnprae numeric
@attribute nnpost numeric
@attribute propn numeric
@attribute adjen numeric
@attribute freq numeric
@attribute neo {0,1}
@attribute loanw {0,1}
@attribute comp {0,1}
@attribute domain {F,K,M,P,T,u}
@attribute medium {B,G,P,S}
@attribute region {MO,MS,MW,NO,NW,SO,SW,UR}
@attribute marker
{ens,es,hochs,ns,s,ses,shoch,0}
@data
0,0,1,1,0,0,0,0,29,0,0,0,F,B,SW,0
0,1,0,1,0,0,0,0,7,0,0,0,F,B,SO,es
0,0,1,1,0,0,0,0,19,0,0,0,K,S,SW,s
0,1,0,1,0,0,0,0,29,0,0,0,P,G,UR,ens
0,0,1,1,0,1,0,0,22,1,0,1,P,G,SW,s
...
```

Figure 6: WEKA import with 16 LMT attributes and sample data

[4] For the noun frequency classification, we used the *DEREWO* ranking lists available at *http://www.ids-mannheim.de/kl/projekte/methoden/derewo.html*.

[5] A decision tree mapping our overall observations is available at *http://hypermedia.ids-mannheim.de/treeText.c095.m2000.pdf.*

[6] For legal reasons – the underlying language corpora contain copyrighted material – users need to register before downloading *GenitivDB* at *http://hypermedia.ids-mannheim.de/call/public/korpus genitivdb.*

[7] In fact all sentences come from the *mk2* subcorpus, which is to some extent balanced with respect to text type.

The evaluation of our model on the training data gives us 92.2% correctly classified gold standard instances (see the summary in figure 7). Considering the fact that the used parameter list excludes some potentially relevant content like phonetic information (stress, vowel length, etc.), this seems both promising and improvable. We will continue experimenting with different attribute selections, including the above mentioned information.

The algorithm performance with precision, recall, and F-scores of more than 90% can be confirmed on the complete *GenitivDB* dataset if we consider only candidates with more than one score point (i.e., reliable genitive nouns whith classified endings based on our distribution rules). Including also the "weak" candidates (i.e., genitive nouns with only one score point) gives us precision values of almost 80%. Again, this seems reasonable, since we expect our automatically generated collection to still contain some incorrect findings, especially among the weak candidates and nouns with zero-marker. The LMT classification errors, together with our score points, serve as valuable starting points for further investigation and corrections (of genitive probability and/or assigned genitive ending) that will be included in future releases of *GenitivDB*.

```
Correctly Classified Instances         271               92.1769 %
Incorrectly Classified Instances        23                7.8231 %
Kappa statistic                          0.8537
Mean absolute error                      0.0238
Root mean squared error                  0.1087
Relative absolute error                 17.4154 %
Root relative squared error             41.9279 %
Total Number of Instances              294

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0          0         0         0          ?        ens
                 0.9       0.057      0.891     0.9       0.896      0.986    es
                 0         0          0         0         0          ?        hochs
                 0         0          0         0         0          ?        ns
                 0.942     0.09       0.936     0.942     0.939      0.987    s
                 1         0          1         1         1          1        ses
                 1         0          1         1         1          1        shoch
                 0.85      0.004      0.944     0.85      0.895      0.999    0
Weighted Avg.    0.922     0.072      0.922     0.922     0.922      0.988

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   <-- classified as
   0   0   0   0   0   0   0   0 |   a = ens
   0  90   0   0   9   0   0   1 |   b = es
   0   0   0   0   0   0   0   0 |   c = hochs
   0   0   0   0   0   0   0   0 |   d = ns
   0  10   0   0 162   0   0   0 |   e = s
   0   0   0   0   0   1   0   0 |   f = ses
   0   0   0   0   0   0   1   0 |   g = shoch
   0   1   0   0   2   0   0  17 |   h = 0
```

Figure 7: Summary of the gold standard evaluation

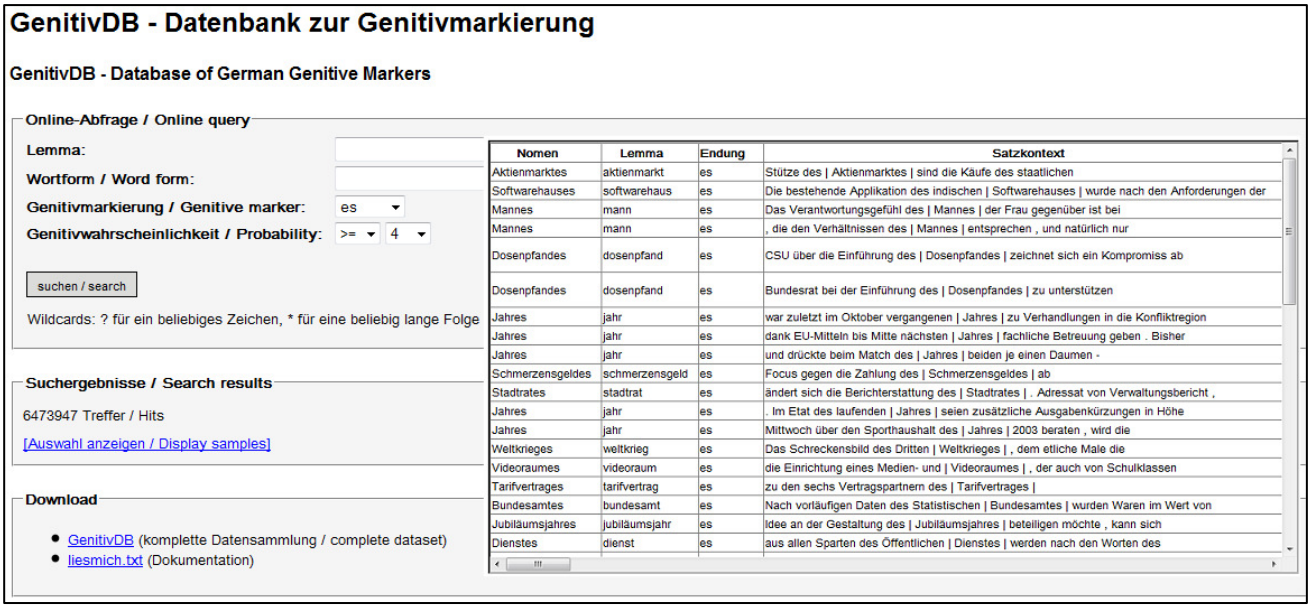

Figure 8: Online retrieval and download of *GenitivDB*

5. Summary and Outlook

We presented an empirical approach to work with large annotated corpora for the explanation of linguistic phenomena, using the example of German genitive markers. The output is a comprehensive NLP resource that – to the best of our knowledge – is unique for contemporary research on German genitive formation. It allows for statistical analysis on a large set of intra- und extralinguistic metadata. Alongside with an online query form and a complete download in CSV format, several data subsets in RDATA (R Workspace File) format will be available in the near future.

Within a pilot study, we examined machine learning algorithms to reveal the influence of factors predicting genitive marking. An elaborated paper on the effective directions and effect sizes of the factors, using established measures like odds ratio and Cramér's V, is underway (Konopka, 2014).

Some of the included factors are possibly interrelated (e.g., frequency class and number of syllables or frequency class and neologism/loanword attributes), so one of our future objectives is to inspect especially the (empirically observable) interrelationships of these factors. In addition, we plan to extend our rather small gold standard collection that nevertheless served well for the evaluation of a prototypical predictive model.

6. References

Appel, E. (1941). *Vom Fehlen des Genitiv-s*. München: Beck.

Baayen, R.H., Piepenbrock, R., and Gulikers, L. (1995). *The CELEX Lexical Database.* CD-ROM. Philadelphia.

Cohen, A. (1980). On the graphical display of the significant components in a two-way contingency table. In *Communications in Statistics - Theory and Methods*, A9, pp. 1025--1041.

Dudenredaktion (2007). *Der Duden in 12 Bänden. Das Standardwerk zur deutschen Sprache: Duden 09. Richtiges und gutes Deutsch: Wörterbuch der sprachlichen Zweifelsfälle*. Mannheim: Bibliographisches Institut.

Dudenredaktion (2009). *Duden 04. Die Grammatik: Unentbehrlich für richtiges Deutsch*. Mannheim: Bibliographisches Institut.

Fehringer, C. (2011). Allomorphy in the German genitive. A paradigmatic account. In *Zeitschrift für Germanistische Linguistik*, 39/1, pp. 90--112.

Hansen, S., Schneider, R. (2013). Decision Tree-Based Evaluation of Genitive Classification - An Empirical Study on CMC and Text Corpora. In Gurevych, I., Biermann, C., Zesch, T. (Eds.), *Language Processing and Knowledge in the Web*. Berlin/Heidelberg: Springer, pp. 83—88.

Konopka, M. (2014). Variation der starken Genitivmarkierung. In *grammis - Korpusgrammatik. http://www.ids-mannheim.de/kogra/.*

Landwehr, N., Hall, M., and Eibe, F. (2005). Logistic Model Trees. In *Machine Learning*, 59, pp. 161–205.

Meyer, D., Zeileis, A., and Hornik, K. (2005). The strucplot framework: Visualizing multi-way contingency tables with vcd. Report 22, Department of Statistics and Mathematics, Wirtschaftsuniversität Wien, Research Report Series.

Pfeffer, J. A., Morrison, S. E. (1984). The genitive singular with -s and/or -es in spoken and written German. In Pfeffer, J. A. (Ed.), *Studies in descriptive German grammar*. Heidelberg: Groos, pp. 9--18.

Schneider, R. (2012). Evaluating DBMS-Based Access Strategies to Very Large Multi-Layer Annotated Corpora. In Calzolari, N. (Ed.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012) Workshops*. Istanbul: European Language Resources Association.

Szczepaniak, R. (2010). Während des Flug(e)s/des Ausflug(e)s? German Short and Long Genitive Endings between Norm and Variation. In Lenz, A. N., Plewnia, A. (Eds.), *Grammar between Norm and Variation*. Frankfurt am Main: Peter Lang, pp. 103—126.

Witten, I., Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco: Morgan Kaufmann.