Investigating the emergence of noun capitalization in a corpus of handwritten texts

Lisa Dücker, Stefan Hartmann and Renata Szczepaniak (University of Hamburg, Germany)

Sentence-internal capitalization of nouns is a distinctive characteristic of the German spelling system. Its diachronic emergence in the 16th and 17th centuries has received much attention in recent decades (e.g. Kämpfert 1980, Moulin 1990). Most importantly, a fairly comprehensive corpus study on the basis of printed texts has contributed significantly to our understanding of this phenomenon (Bergmann & Nerius 1998). However, some important questions regarding the emergence of sentence-internal capitalization remain unanswered. Firstly, previous studies have only provided monofactorial analyses, focusing on a small number of semantic and pragmatic factors such as animacy and reverence. The interaction of these different factors, as well as additional ones like word frequency, has not been taken into account so far. Secondly, previous research has largely focused exclusively on printed texts (with the exception of Moulin 1990, who only investigates the writings of one author). However, it seems promising to take handwritten texts into account as well, as they are produced in a more spontaneous way as compared to printed texts and can therefore provide a glimpse into (quasi-) spontaneous language production. Handwritten texts might therefore be particularly well-suited for investigating the impact of cognitive, semantic, and syntactic factors in the use of sentence-internal capitalization.

In order to investigate the emergence of capitalization in Early New High German handwritten texts in more detail, we compiled a corpus of witch interrogation protocols, based on the edition by Macha et al. (2005), which comprises text samples from six dialect areas and covers the time span from 1570 to 1653. This period marks the incipient stages of sentence-internal capitalization. Consequently, there is a large amount of variation in the texts, which allows for investigating the factors that boost the use of sentence-internal capital letters. Fig. 1 shows the regional distribution of the texts. The six different shapes used for the individual datapoints represent the language areas assumed by Macha et al. Using digital facsimiles, we double-checked for samples of each text that the edition faithfully represents the handwritten originals. In sum, the corpus contains about 90,000 tokens. A multi-layer annotation was implemented, sentence boundaries were tagged according to a complex set of criteria (as punctuation is sparse and unreliable in many of the texts; cf. Szczepaniak & Barteld 2016 for details), and all tokens tagged as nouns were annotated for animacy using a fine-grained animacy hierarchy, as established hierarchies like the one proposed by Zaenen et al. (2004) proved inadequate for the given text type (cf. Barteld, Hartmann & Szczepaniak 2016).

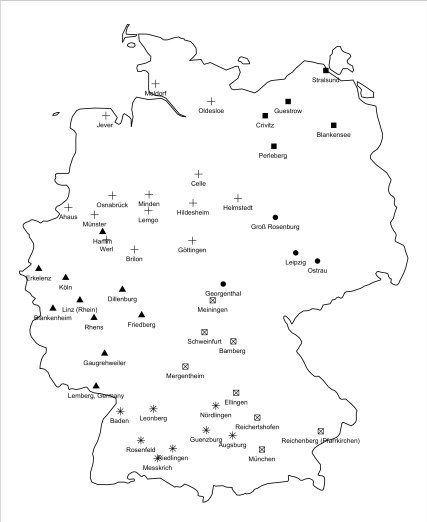

Fig. 1: Regional distribution of the corpus texts.

Drawing on these data, we fit a binomial mixed-effects regression model using R (R Core Team 2016) and *lme4* (Bates et al. 2015). Two variables were entered as fixed effects into the model: Animacy and Frequency. The fine-grained animacy annotation was broken down to a five-way distinction ("abstract", "concrete", "animal", "human", "superhuman", the latter including terms like *God, devil, demon*, which are fairly frequent in the given text type), as a too fine-grained variable would lead to a rank-deficient model. The idea behind adding frequency as a factor is that we can expect more frequent words to have a fairly fix graphemic *gestalt*, while we would expect more variation for low-frequency items (cf. e.g. Kapatsinski 2010). The token frequency of each type was obtained from the corpus itself. An alternative would have been to use another, larger corpus, but we decided against this option for two reasons: Firstly, there is currently no significantly larger corpus covering the time period in question, and secondly, the vocabulary used in the witch interrogation protocols is partly very specific to this text type. In addition, random intercepts for Lemma and Protocol as well as random slopes for the fixed effect of Animacy were added. Tab. 1 shows the results of the model.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.01	0.25	-8.05	<0.001***
Animacy-concrete	1.27	0.27	4.62	<0.001***
Animacy-animal	1.7	0.59	2.87	<0.001***
Animacy-human	1.82	0.38	4.81	<0.001***
Animacy-superhuman	3.79	1.28	2.97	<0.001***
$\log_{10}$(Freq)	0.39	0.2	1.9	0.06 .
Animacy-conc×$\log_{10}$(Freq)	-0.85	0.26	-3.32	<0.001***
Animacy-anim×$\log_{10}$(Freq)	-0.17	0.68	-0.25	0.8
Animacy-hum×$\log_{10}$(Freq)	-0.35	0.32	-1.11	0.27
Animacy-sup×$\log_{10}$(Freq)	-1.79	0.76	-2.36	0.02 *

Tab. 1: Coefficients of the fixed effects in the mixed model.

According to a log-likelihood test, the use of the random intercepts and slopes mentioned above is warranted as the model performs significantly better than models without the random effects for Lemma and Protocol. Importantly, the model also performs better than null models without Animacy (χ^2=327, df=22, p<0.001***) or without the interaction term between Animacy and Frequency (χ^2=15.7, df=4, p=0.003**). The index of concordance *C*, which assesses how well the model predicts the data (Baayen 2008: 204), indicates a very good fit (C=0.94, considered "outstanding discrimination" by Hosmer & Lemeshow 2000: 162). All Variance Inflation Factors (VIFs) – which are used to check for potential multicollinearity – are below the threshold of 5, which is often mentioned as a rule of thumb in the statistical literature (Levshina 2015: 272). In sum, the present study confirms the results which Barteld et al. (2016) have obtained on the basis of a much smaller and less representative sample: While the effect of animacy shown in earlier studies is very clearly substantiated, animacy alone cannot explain the variation in the data. We predicted that token frequency might have an effect on capitalization, but according to the results of the model, this effect is not straightforward. In general, highly frequent items seem more prone towards capitalization than low-frequency items. However, there is much lexeme-specific variation among high-frequency items which can partly

be explained in terms of animacy. For instance, *Mann* 'man' is capitalized in 69 out of 119 cases (81%) and *Gott* 'god' in 174 out of 189 (92%), while only 18 out of 117 occurrences of *Tag* 'day' (15%) are capitalized. However, some possibly important factors have not been captured in the model yet. Aspects that need further investigation include pragmatic motivations and syntactic factors in the early use of sentence-internal capitalization. For example, it is striking that in a sample of 18 texts, terms denoting women are significantly less often capitalized than terms denoting men (Fisher exact test: $p<0.001$, odds ratio=3.01), which points to an influence of cultural factors. In addition, future research could extend the focus from bare nouns to entire NPs, taking their internal structure into account.

Acknowledgments

The research reported on in this paper was funded by the German Research Foundation (2013–2015: SZ 280/2-1, KO 909/12-1; 2016–2018: SZ 280/2-3).

References

Baayen, R. H. (2008). Analyzing Linguistic Data. A Practical Introduction to Statistics using R. Cambridge: Cambridge University Press.

Barteld, F., Hartmann, S., & Szczepaniak, R. (2016). The usage and spread of sentence-internal capitalization in Early New High German: A multifactorial approach. *Folia Linguistica*, *50*(2), 385–412.

Szczepaniak, R., & Barteld, F. (2016). Hexenverhörprotokolle als sprachhistorisches Korpus. In S. Kwekkeboom & S. Waldenberger (Eds.), *PerspektivWechsel oder: Die Wiederentdeckung der Philologie. Bd. 1 Sprachdaten und Grundlagenforschung in Historischer Linguistik* (pp. 43–70). Berlin: Erich Schmidt Verlag.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using *lme4*. Journal of Statistical Software, 67(1). http://doi.org/10.18637/jss.v067.i01

Bergmann, R., & Nerius, D. (1998). Die Entwicklung der Großschreibung im Deutschen von 1500-1700. 2 Bde. Heidelberg: Winter.

Hosmer, D. W., & Lemeshow, S. (2000). Applied Logistic Regression (2nd ed.). New York: Wiley.

Kämpfert, M. (1980). Motive der Substantivgroßschreibung. Beobachtungen an Drucken des 16. Jahrhunderts., 99, 72–98.

Kapatsinski, V. (2010). What is it I am writing? Lexical frequency effects in spelling Russian prefixes: Uncertainty and competition in an apparently regular system. Corpus Linguistics and Linguistic Theory, 6(2), 157–215.

Macha, J., Topalović, E., Hille, I., Nolting, U., & Wilke, A. (Eds.). (2005). Deutsche Kanzleisprache in Hexenverhörprotokollen der Frühen Neuzeit. Bd. 1: Auswahledition. Berlin, New York: De Gruyter.

Moulin, C. (1990). Der Majuskelgebrauch in Luthers deutschen Briefen: (1517 - 1546). Heidelberg: Winter.

R Core Team. (2016). R. A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing.

Zaenen, A., Carletta, J., Garretson, G., Bresnan, J., Koontz-Garboden, A., Nikitina, T., … Wasow, T. (2004). Animacy Encoding in English: Why and How. In B. Webber

& D. Byron (Eds.), DiscAnnotation ’04 (pp. 118–125). Stroudsburg, PA: Association for Computational Linguistics.