Frequencies of Nouns, Pronouns, and Verbs by Document Types in German and Spanish Web Corpora

In this exploratory study, we analyze the relative frequencies of nouns, pronouns, and verbs in random samples of documents from two of the COW web corpora (Schäfer & Bildhauer, 2012): the German DECOW2012 (9.1 billion tokens from the *.de* top-level domain) and the Spanish ESCOW2012 (1.6 billion tokens from the *.es* top-level domain). More specifically, we look for correlations of these frequencies with the categories of a simple genre/register classification based on Sharoff (2006), which is itself EAGLES-based, and which distinguishes between *Authorship* (5 categories), *Mode* (4 categories), (intended) *Audience* (3 categories), *Aim* (5 categories), *Domain* (8 categories). In order to be able to classify blog and forum documents appropriately, we added a category for *Mode*, namely *Quasi-Spontaneous* for unmoderated forum discussions and similar documents. In a two-rater evaluation of the categorization scheme on a DECOW2012 sample, we reached an inter-rater agreement (Cohen's Kappa) of $\kappa>0.9$ for *Mode*, $\kappa>0.8$ for *Authorship* and *Domain*, and $\kappa>0.6$ for *Aim* and *Audience*. To derive comparable frequencies of the relevant parts of speech in the samples used here, we mapped the STTS tagset and the Spanish TreeTagger tagset to a simple common tagset comprising 14 tags.

A number of interesting correlations emerges, which can be used to characterize the respective segments of the Web. Moreover, we explore the potential of these features in automatic document classification. While similar features have been used in multivariate models for classifying documents written in standard English (starting with Karlgren and Cutting's (1994) work on the Brown corpus and applied to Web corpora in, e.g., Sharoff, 2010), our main focus is on distinguishing between the more informal language used in the *Quasi-Spontaneous* mode on the one hand, and the traditional *Written* mode.

In a binary logistic regression on the response variable “document is classified as *Quasi-Spontaneous* (1) or not (0)“, only the relative noun frequency comes up as significant (but highly significant) in both languages. Furthermore, the frequencies of nouns and pronouns varies significantly between major *Aim* categories in German. To illustrate, using the arithmetic means (weighted by document size): *Discussion* has $x_{Noun}=0.18$ and $x_{Pronoun}=0.11$, but *Information* has $x_{Noun}=0.24$ and $x_{Pronoun}=0.06$. Interestingly, these contrasts are much weaker (non-significant) in Spanish. Both in German and Spanish, no significant correlation can be observed between the *Audience* classification (*General*, *Informed*, *Professional*) and POS frequencies. The relative frequencies in the eightfold *Domain* classification also vary significantly (e.g., above average noun frequencies and below average pronoun frequencies in scientific texts, etc.), although the overall variance is quite high.

In the discussion, we sketch further evaluations of these results, trying to pinpoint the influences of actual language-specific factors, different distributions of document types under the relevant top-level domains, and even POS tagset selection and tagger error rate (surprisingly not taken into account in, e.g., Sharoff, 2010) in the different document types, considering that some categories (like *Quasi-Spontaneous*) imply to a large extend a greater openness towards substandard language, and that tagger accuracy on web corpora is generally lowered (Giesebrecht & Evert, 2009).

References

Giesbrecht, E. & Evert, S. (2009), Part-of-Speech (POS) Tagging – a Solved Task? An Evaluation of POS Taggers for the German Web as Corpus, *in* Iñaki Alegria; Igor Leturia & Serge Sharoff, ed., 'Proceedings of the Fifth Web as Corpus Workshop (WAC5)', pp. 27–35.

Karlgren, J. & Cutting, D. (1994), Recognizing Text Genres with Simple Metrics Using Discriminant Analysis, *in* 'Proceedings of the 15th conference on Computational linguistics', pp. 1071–1075.

Lee, D. (2001), 'Genres, registers, text types, domains, and styles: Claryfying the concepts and navigating a path through the BNC jungle', *Language Learning and Technology* 5(3), 37–72.

Schäfer, R. & Bildhauer, F. (2012), Building Large Corpora from the Web Using a New Efficient Tool Chain, *in* Nicoletta Calzolari et al., ed., 'Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)', ELRA, Istanbul, pp. 486–493.

Sharoff, S. (2006), Creating General-Purpose Corpora Using Automated Search Engine Queries, *in* Marco Baroni & Silvia Bernardini, ed., 'WaCky! Working papers on the Web as Corpus', GEDIT, Bologna.

— (2010), In the garden and in the jungle: comparing genres in the BNC and internet, *in* Alexander Mehler; Serge Sharoff & Marina Santini, ed., 'Genres on the Web', Springer, Heidelberg etc., pp. 149–166.