Operationalizing formality in European Parliament data: Bringing together human judgements and corpus-driven detection

Ilmari Ivaska (University of Turku) & Adriano Ferraresi (University of Bologna)
keywords: formality; corpus-driven methods; Universal Dependencies; European Parliament

Formality differences are often considered central in describing and explaining linguistic variation (e.g. in multidimensional analyses, see Biber 1989 and all the subsequent work). Formality serves as an explanatory tool when interpreting the differences across texts categories (e.g. registers or genres). Suggested measures of formality have often been tested in terms of their capability to account for category variation (as pointed out by Li et al. 2016), sometimes without any objective grounding in human perception of formality. In this paper, we explore a data-driven and human-informed operationalization of linguistic formality for European Parliament data. We take as our point-of-departure two text classes that we hypothesize to be characterized by different levels of formality, i.e. speeches delivered impromptу vs. read-out, establish their order of perceived formality by means of human judgements, and use that as a gold standard in a data-driven analysis of linguistic formality differences between the said text classes. Our specific research questions are: 1) Are read-out speeches perceived as more formal than impromptu speeches? 2) Which linguistic features contribute to distinguishing these text classes?

Our data come from the EPTIC corpus (Bernardini, Ferraresi & Miličević 2016), where we use the English native speaker subset. We include texts that have originally been either read-out or delivered impromptu and split these subsets randomly to train (80%) and test (20%) sets. Contrary to much of earlier work on formality (e.g. Graesser et al. 2014), we are interested in identifying the linguistic constructions that contribute to the difference rather than classifying the texts. Thus, we use the train set to detect the most consistent linguistic differences between the read and impromptu texts. The data-driven analysis consists of 3 phases:

1) Annotate the data using a neural parser that complies with the Universal Dependencies annotation scheme (Kanerva et al. 2018);
2) Extract the frequencies of syntactically defined part-of-speech bigrams (e.g. NOUNNODE_nsubj_VERBHEAD for *man–writes*) and use them as our feature set;
3) Use the Boruta feature selection (Kursa & Rudnicki 2010) to detect those bigrams whose frequency consistently distinguishes read-out and impromptu texts;

As pointed out by Ivaska & Bernardini (submitted), such bigrams successfully capture multiple linguistic phenomena (POS, syntactic functions, word order and hierachical relations) and are still qualitatively interpretable.

We use the test set in a human judgement survey conducted online. In the survey, read-out and impromptu texts are paired and the participants are asked to choose the more formal of the two. The text pairings are randomly assigned respondents; each responent is asked to evaluate 10 text pairs and each text is evaluated at least 10 times. Our gold standard consists of texts that are consistently evaluated as either more or less formal than their respective paired texts. Finally, we train a Random Forest prediction model with the detected POS bigrams on the train data to see, how well they predict the test data. The performance of the prediction model is used to validate the reliability of the detected features as formality indicators in the EP context.

word count: 497

References

Bernardini, Silvia, Adriano Ferraresi & Maja Miličević. 2016. From EPIC to EPTIC — Exploring simplification in interpreting and translation from an intermodal perspective. *Target* 28(1). 61–86.

Biber, Douglas. 1989. A typology of English texts. *Linguistics* 27(1). 3–43.

Graesser, Arthur C., Danielle S. McNamara, Zhiqiang Cai, Mark Conley, Haiying Li & James W. Pennebaker. 2014. Coh-Metrix measures text characteristics at multiple levels of language and discourse. *Elementary School Journal* 115(2). 210–229.

Ivaska, Ilmari & Silvia Bernardini. submitted. Multi-varietal corpus research and the paradox of comparability: How can we cope with the unavoidable?

Kanerva, Jenna, Filip Ginter, Niko Miekka, Akseli Leino & Tapio Salakoski. 2018. Turku Neural Parser Pipeline: An End-to-End System for the CoNLL 2018 Shared Task. *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Brussels, Belgium: Association for Computational Linguistics.

Kursa, Miron & Witold Rudnicki. 2010. Feature Selection with the Boruta Package. *Journal of Statistical Software, Articles* 36(11). 1–13.

Li, Haiying, Arthur C. Graesser, Mark Conley, Zhiqiang Cai, Phillip I. Pavlik jr & James W. Pennebaker. 2016. A new measure of text formality: An analysis of discourse of Mao Zedong. *Discourse Processes* 53. 205–232.