"Models - predictions - data: an (un)problematic relationship? The example of Systemic Functional Linguistics (SFL)"

Erich Steiner

There is significant gap between the high level of abstraction of linguistic models, such as SFL (Fawcett 2006, Halliday and Matthiessen 2014), and data provided through shallow analysis and annotation of electronic corpora (Alves et al 2010, Steiner 2012, Hansen-Schirra, Neumann and Steiner 2012, Kunz et al. 2017). This is one of the reasons why work in my own group has used SFL for generating hypotheses and for interpreting theory-neutral data, but rarely as annotations in the data directly. Direct SFL-annotations are a) costly, b) inconsistent between annotators, and c) make the data theory-dependent.

In a range of different studies using data-mining techniques (Taboada et al 2011, Degaetano-Ortlieb et al 2014), attempts have been made to use ideas from SFL as an underlying linguistic model to some extent. The latter use 3 levels of analysis (shallow features, a limited set of features from register theory, and finally a combination of these). The combination shows improved results for text classification compared to pure "bag-of-words"-type approaches, at least in situations in which linguistic differences of registers are marked. Direct SFL-annotations in data are possible, at least for well-operationalized register features. There is a question of how SFL-specific (some of) these features really are.

In order to further narrow the gap between SFL-theorizing and data, improved strategies have to be developed of formulating empirically-testable hypotheses. Two examples of studies of cohesive chains will be outlined to approach this goal (Kunz et al 2016, Laphinova-Koltunski et al 2016). Yet once more the question is how SFL-specific (some of) these hypotheses and their operationalizations are

Particular challenges for SFL as an underlying model result from the fact that SFL annotations constitute highly interpreted data. They are difficult in terms of inter-coder consistency and expensive to create in sufficient quality. The resulting problems for outside evaluation and repeatability of studies, one important way of enhancing quality of empirical research, need to be addressed.

References:

Alves, Fabio, Adriana Pagano, Stella Neumann, Erich Steiner & Silvia Hansen-Schirra. 2010. "Translation Units and Grammatical Shifts: Towards an Integration of Product- and Process-based Translation Research". In: Shreve, Gregory.M. & Erik Angelone (eds.) *Translation and Cognition*. Amsterdam: John Benjamins.109-142

Degaetano-Ortlieb, S., Fankhauser, P., Kermes, H., Lapshinova-Koltunski, E., Ordan, N. and Teich, E. (2014). Data Mining with Shallow vs. Linguistic Features to Study Diversification of Scientific Registers *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC 2014)*. Reykjavik, Iceland.

De Sutter, Gert, Isabelle Delaere & Marie-Aude Lefer. 2017. 'Empirical Translation Studies. New Theoretical and Methodological Traditions.'. Trends in Linguistics Studies and Monographs. Berlin and New York: Mouton de Gruyter

Fawcett, R.P. 2008. *Invitation to Systemic Functional Linguistics through the Cardiff Grammar*. London: Eqinox

Halliday, M.A.K. and Christian.M.I.M. Matthiessen. 2014. *An Introduction to Functional Grammar* London: Arnold (earlier versions by Halliday in 1985/ 1994, Halliday and Matthiessen 2004).

Hansen-Schirra, Silvia, Stella Neumann and Erich Steiner. 2012. *Cross-linguistic Corpora for the Study of Translations. Insights from the language pair English – German*. Series Text, Translation, Computational Processing. Berlin, New York: Mouton de Gruyter

Kerstin Kunz, Stefania Degaetano-Ortlieb, Ekaterina Lapshinova-Koltunski; Katrin Menzel and Erich Steiner. 2017. "English-German contrasts in cohesion and implications for translation". in: De Sutter et al. 2017

Kunz, K., E. Lapshinova-Koltunski and J.M. Martinez-Martinez (2016). Beyond Identity Coreference: Contrasting Indicators of Textual Coherence in English and German. In Proceedings of CORBON at NAACL-HLT2016, San Diego, 16 June.

Lapshinova-Koltunski, E., J. Martinez Martinez, K. Kunz, E. Steiner and K. Menzel (2016). Lexical Cohesion: Dimensions and Linguistic Properties of Chains in English and German. ICAME-37, Hong Kong.

Steiner, Erich. 2012. "Methodological cross-fertilization: empirical methodologies in (computational) linguistics and translation studies", in: *Translation: Computation, Corpora, Cognition*. Special Issue "At the Crossroads between Contrastive Linguistics, Translation Studies and Machine Translation." TC3 website (http://www.t-c3.org). Vol.2 No.1. 2012. pp 3-21.

Taboada, M., J. Brooke, M. Tofiloski, K. Voll and M. Stede (2011) Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics* 37 (2): 267-307