Axel Wisiorek

Quantitative Methods of a Cognitive Text Typology: Automatic Genre Classification as a Reconstruction of Cognitive World Models
DOI: https://doi.org/10.5282/oph.19

English Summary

This study proceeds from the basic cognitive-linguistic assumption that the structuring of texts as sequences of speech acts follows genre-specific rules and patterns forming part of the cognitive knowledge system of a speaker within a language community, namely in the form of schematic structural models (text world models, Schulze 2018; 2019; 2020). Learned through cognitive processes of conventionalization and classification via the typification of recurrent structural patterns in language use, these abstract, usage-based cognitive models (Johnson-Laird 1983; Langacker 2000) regulate the production and reception of texts of a given genre (Miller 1984). This involves adapting the structure of the individual cognitive text models constructed during text processing (situation models, van Dijk & Kintsch 1983) to the specific type of communicative situation in which the texts arise.

When applied as cognitive structural rules in text production, the genre-specific schemata laid down in a text world model establish a trace (Langacker 2000; Schwarz 2000) within the linguistic structure of a text. This study will explore statistical classification methods for identifying such genre-typical text structure patterns (Heinemann & Viehweger 1991; Fix 2008) based on recurring, quantitative features in texts of a given genre. The text grammars (van Dijk 1972) resulting from the extraction of prototypical structural feature values from corpora as language usage data can then ultimately facilitate the reconstruction of the schematic rules of the corresponding text world models by means of interpreting their relational, referential and information-structural characteristics as parameters of a cognitive text linguistics (Schulze 2018; 2019; 2020).

For this intended corpus-based exploration of cognitive genre models, this study relies on quantitative text structure parameters such as information density, elaboration measures or frequent event patterns that can be assumed to be relevant to the construction of cognitive text models, with their selection being primarily based on the taxonomy of parameters of text world models (TWM parameters) as developed by Schulze (2018; 2019; 2020). To operationalize these, feature construction methods of representing texts via both multivariate text-structural feature sets and sequence-based text structure patterns are introduced, including their extraction from annotated text corpora using data mining techniques. Appropriate clustering and classification methods for the exploratory study of such text structure representations are proposed, including Random Forest classifiers and Dynamic Time Warping-based clustering, which eventually allow for a quantitative cognitive text typology based on TWM parameters.

In a concluding pre-test study for such a cognitively grounded genre classification, the established methods for representing and classifying texts based on TWM parameters as linguistically encoded, schematic patterns for constructing genre-specific cognitive text models are tested on an information-structurally annotated, historically and dialectally stratified corpus of Ob-Ugrian folk tales as well as texts from other genres. Since these texts of oral tradition are close to the original language practice (Schulze 2019), they are well suited as test cases for the intended usage-based identification of genres and subgenres as text structure pattern types which can be interpreted as culture-specific orders of discourse (Foucault 1981; 1991; Schulze 2020). As the study shows, several such types of structural organization can indeed be recognized in the Ob-Ugric corpus via the proposed TWM operationalization.