

Description of PhD. Project Using Genre Analysis for Constructing an Ontology of US Patents

The data used for the project will be two versions of a corpus of US patents of approximately 7 million words. One version of the corpus will be cleaned of “noise” data that may distort the statistical analysis. The other version of the corpus will be left untouched for a qualitative analysis of the corpus (distribution of sections, use of mathematical formulas, items ordered with bullets, numbers or other, etc). The tools currently available for the corpus analysis are the following: Antconc 3.2.1 (which allows regular expression queries), R.J.C. Watt's Concordance 3.2, WordSmith Tools 4.0 (which has several specific statistical analysis tools), and Corpus Presenter (which has Part-of-Speech tagging tools). The corpus data will be analysed with a methodological framework that combines Swales (1990), Bhatia (1993), and Paltridge (1997) Genre Analysis with Temmerman (2000) and Temmerman & Kerremans (2003) Termontography. Swales and Bhatia established an influential model for identifying the conventions of discourse communities within the communication genres of those discourse communities. Essentially, they have shown how the rhetorical structure of a professional genre links the specific use of language to the specific communication expectations of the discourse community that uses the genre. Paltridge incorporated into this model Fillmore's frame semantics, being of particular importance the concept of prototypicality which would account for variability within a genre. Temmerman & Kerremans derive their methodology for extracting knowledge towards the construction of ontologies from the same idea of prototypicality, thus this project considers that prototypicality can occur at different levels (words, sections, genres) so the information obtained from a genre analysis can be transformed by means of termontographic methods into an ontology of a genre. One of the aspects perceived in the conception of this project is that such an ontology would be very much influenced by its intended receivers. It seems convenient to develop an ontology for receivers interested in extracting information from patents rather than those interested in writing them. The reason is that in genre analysis the traditional approach would be to work for people who had to write a specific genre, discarding options that statistically were less typical. The application of this work to receivers who want to extract information would be a relatively novel approach.

My Spanish Ph.D. supervisors have suggested that I use the software Protégé to build the patent ontology.

REFERENCES

- Bhatia, Vijai K.* (1993). *Analysing Genre: Language Use in Professional Settings*. Longman
- Paltridge, Brian* (1997). *Genre, Frames and Writing in Research Settings*. John Benjamins.
- Swales, John M.* (1990). *Genre Analysis*. Cambridge University Press.
- Temmerman, Rita* (2000). *Towards New Ways of Terminology Description: The Sociocognitive Approach*. John Benjamins.
- Temmerman, Rita & Kerremans, Koen* (2003). *Termontography: Ontology Building and the Sociocognitive Approach to Terminology Description*. In Hajičová, E., Kotěšovcová, A., Mírovský, J. (eds.), *Proceedings of CIL17*, Matfyzpress, MFF UK (CD-ROM). Prague, Czech Republic. Stable URL: http://www.ffpoirot.org/Publications/temmerman_art_prague03.pdf
Accessed: 30/10/2008