Extracting Linguistic Terminology from Scientific Corpora

Christian Lang, Roman Schneider and Karolina Suchowolec (IDS Mannheim)

The compilation of terminological vocabularies plays a central role in the organization and retrieval of scientific texts. Both simple keyword lists as well as sophisticated modellings of relationships between terminological concepts can make a most valuable contribution to the analysis, classification, and finding of appropriate digital documents, either on the web or within local repositories. This seems especially true for long-established scientific fields with various theoretical and historical branches, where the use of terminology within documents from different origins is sometimes far from being consistent.

However, the manual compilation and organization of key terms for a certain domain is time consuming and bound to be subjective. That is why recent developments in the context of Natural Language Processing (NLP) and Digital Humanities (DH) automate the finding – and sometimes even the rough classification – through statistical and linguistic methods. Against this background, we present a novel approach for the computation of a terminological knowledge base that serves as a robust backend for the scientific exploration of linguistic resources, with a focus on grammatical content. Both use case and data basis is the grammatical online system grammis, hosted by the Institute for German Language (IDS) in Mannheim. Grammis is a specialist hypertext resource that brings together terminological, lexicographical, and bibliographic information about German grammar. Initiated more than two decades ago, it combines traditional description of grammatical structures with the results of corpus-based studies. From a technical point of view, all primary data and meta data is coded within more than one thousand semi-structured XML instances that are composed of semantical markup element types ("title", "header", "example", "link anchor", etc.).

Given the broad scope of grammis and the need to continuously update the database, we use automatic key term extraction techniques to provide content for our terminology management system. We are currently investigating three well-known and well-established statistical methods of key term extraction: TF-IDF (term frequency, inverse document frequency, Sparck Jones, 1972), Gries DP (Gries, 2008), and Weirdness Ratio (Gillam et al., 2008). As a plus, we exploit the semantic markup of heterogeneous text sections, coded with XML element types. In a preliminary study, we use different corpus stratifications (specialized vs general corpora based on DEREKO) and compare recall and precision performance of n-gram extractions against a human annotated gold standard.

The extracted grammar terminology is currently managed within different tools, depending on whether it is used in the grammis dictionaries, ontology, or bibliography. As of now, this terminology management is being re-designed and integrated to accommodate terminology from different research projects as well as needs of heterogeneous user groups. The key design principle for the new system is to serve as both – concise grammar reference and a repository for enhancing information retrieval. The system needs then to combine the aforementioned resources into one tool; therefore, it should be able to manage not only descriptions of isolated concepts and terms, but also hierarchical, associative, and equivalency relations between them. For the sake of transparency and open access to content, we prefer a solution that supports standard exchange formats such as TBX by ISO, and SKOS by W3C.

References

Gillam, Lee; Mariam Tariq, Kurshid Ahmad (2005): "Terminology and the construction of ontology". Terminology 11(1): 55–81.

Gries, Stefan Th. (2008): "Dispersions and adjusted frequencies in corpora". International Journal of Corpus Linguistics 13(4): 403–437.

ISO 30042:2008-12 (E) (2008): "Systems to manage terminology, knowledge and content – TermBase eXchange (TBX)".

Kupietz, Marc; Holger Keibel (2009): "The Mannheim German Reference Corpus (DEREKO) as a basis for empirical linguistic research". Makoto Minegishi, Yuji Kawaguchi (Eds.): Working Papers in Corpus-based Linguistics and Language Education, No. 3. Tokyo:

Tokyo University of Foreign Studies (TUFS), 53–59. http://cblle.tufs.ac.jp/assets/files/publications/working_papers_03/section/053-059.pdf [201606-16].

Sparck Jones, Karen (1972): “A statistical interpretation of term specificity and its application in retrieval”. Journal of Documentation 28(1): 11–21.

W3C (2009): “SKOS Simple Knowledge Organization System Reference”. W3C Recommendation 18 August 2009. https://www.w3.org/TR/2009/REC-skos-reference-20090818 [2016-06-16].