Automatic Extraction of the Phraseology of a Legal Subdomain

Fabienne Fritzinger, Ulrich Heid & Nadine Siegmund
Universität Stuttgart, Germany

This paper deals with the identification of phraseological word pairs and multiword groups from corpora of German juridical texts. We intend to answer the following methodological and technical research questions: if we use simple frequency-based extractors on juridical texts from different domains, can these provide the phraseology of a given juridical subdomain by contrasting multiwords from different domains? And can we identify phraseological sequences longer than two items by systematically analysing the context of word pairs? We also wish to contribute to a more detailed description of the use of adverbs in juridical phraseology.

Jurists writing about juridical topics which fall outside their specialization may be acquainted with the terminology of the “semi-foreign” domain, but they will need dictionary support for the related phraseology. In the framework of preparations for a German juridical dictionary (cf. Heid et al. 2008 for related attempts), we aim at automatically extracting text samples from a large corpus of juridical literature from different domains, to support the detailed lexicographic description of the respective phraseology.

The dictionary should provide information on the following types of collocations: adjective+noun; noun+nominal or prepositional attribute; verb+subject/object noun; verb/adjective+adverb. Moreover, we intend to capture the morphosyntactic behaviour of the collocations (article use, modifiability, number restrictions, lexical variation), and larger phraseological groups, e.g. clusters of verb+adverb+object phrases (e.g. Mangel arglistig verschweigen). Adverb use in the German juridical language is highly specific and – to our knowledge – not well documented.

Our procedures start from single word term candidates of a given legal subdomain (extracted by means of relative frequencies (Ahmad et al. 1992)) which are compared with terms of general language texts and of texts from other subdomains to single out subdomain vocabulary (cf. the schema in Figure 1 for an abstract overview). Based on the most relevant of these candidates, significant word pairs are extracted from syntactically analysed data (dependency parsed, Schiehlen 2003). The use of dependency parsing allows us to identify grammatical relationships (e.g. in verb+object pairs, cf. the screendump in Figure 2), even if they are not adjacent, as it often happens in German.

A detailed analysis of the context of the word pairs provides data about adverb use and about (small) lexical sets of adverbs intervening in e.g. verb+object groups.
Figure 1: Extraction schema for single word terms.

Figure 2: Verb+object pairs, including example sentences: insolvency legislation.

