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. *M¨angel 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.

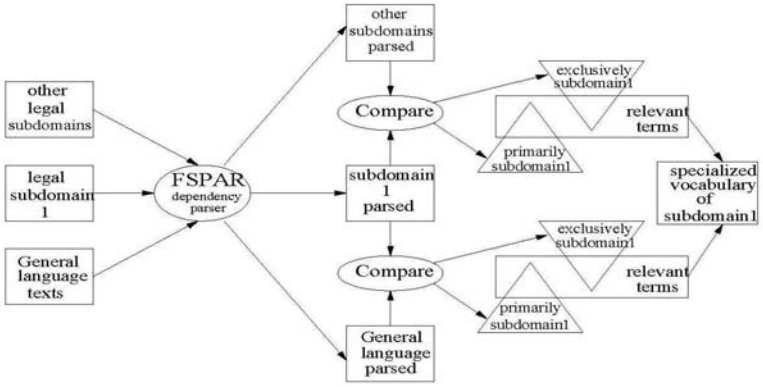
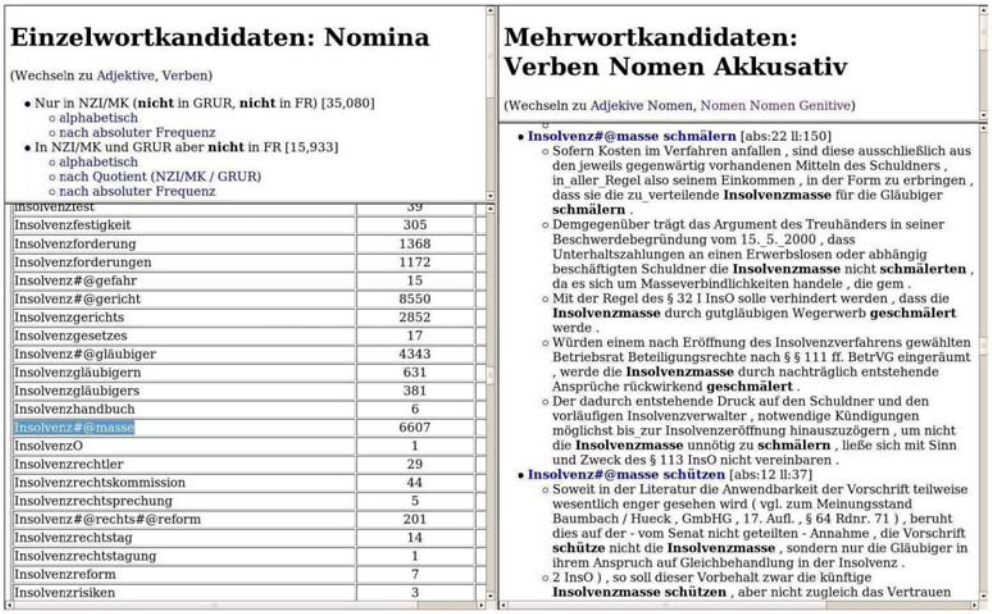


Figure 1: Extraction schema for single word terms.



Einzelwortkandidaten: Nomina

(Wechseln zu Adjektive, Verben)

- Nur in NZI/MK (**nicht** in GRUR, **nicht** in FR) [35,080]
  - alphabetisch
  - nach absoluter Frequenz
- In NZI/MK und GRUR aber **nicht** in FR [15,933]
  - alphabetisch
  - nach Quotient (NZI/MK / GRUR)
  - nach absoluter Frequenz

Insolvenzfest	39
Insolvenzfestigkeit	305
Insolvenzforderung	1368
Insolvenzforderungen	1172
Insolvenz#@gefahr	15
Insolvenz#@gericht	8550
Insolvenzgerichts	2852
Insolvenzgesetzes	17
Insolvenz#@gläubiger	4343
Insolvenzgläubigern	631
Insolvenzgläubigers	381
Insolvenzhandbuch	6
Insolvenz#@masse	6607
InsolvenzO	1
Insolvenzrechtler	29
Insolvenzrechtskommission	44
Insolvenzrechtsprechung	5
Insolvenz#@rechts#@reform	201
Insolvenzrechtstag	14
Insolvenzrechtstagung	1
Insolvenzreform	7
Insolvenzrisiken	3

Mehrwortkandidaten: Verben Nomen Akkusativ

(Wechseln zu Adjekive Nomen, Nomen Nomen Genitive)

- **Insolvenz#@masse schmälern** [abs:22 ll:150]
  - Sofern Kosten im Verfahren anfallen , sind diese ausschließlich aus den jeweils gegenwärtig vorhandenen Mitteln des Schuldners , in_aller_Regel also seinem Einkommen , in der Form zu erbringen , dass sie die zu_verteilende **Insolvenzmasse** für die Gläubiger **schmälern** .
  - Demgegenüber trägt das Argument des Treuhänders in seiner Beschwerdebegründung vom 15._5._2000 , dass Unterhaltszahlungen an einen Erwerbslosen oder abhängig beschäftigten Schuldner die **Insolvenzmasse** nicht **schmälerten** , da es sich um Masseverbindlichkeiten handele , die gem .
  - Mit der Regel des § 32 I InsO solle verhindert werden , dass die **Insolvenzmasse** durch gutgläubigen Wegerwerb **geschmälert** werde .
  - Würden einem nach Eröffnung des Insolvenzverfahrens gewählten Betriebsrat Beteiligungsrechte nach § § 111 ff. BetrVG eingeräumt , werde die **Insolvenzmasse** durch nachträglich entstehende Ansprüche rückwirkend **geschmälert** .
  - Der dadurch entstehende Druck auf den Schuldner und den vorläufigen Insolvenzverwalter , notwendige Kündigungen möglichst bis_zur Insolvenzeröffnung hinauszuzögern , um nicht die **Insolvenzmasse** unnötig zu **schmälern** , ließe sich mit Sinn und Zweck des § 113 InsO nicht vereinbaren .
- **Insolvenz#@masse schützen** [abs:12 ll:37]
  - Soweit in der Literatur die Anwendbarkeit der Vorschrift teilweise wesentlich enger gesehen wird (vgl. zum Meinungsstand Baumbach / Hueck , GmbHG , 17. Aufl. , § 64 Rdnr. 71) , beruht dies auf der - vom Senat nicht geteilten - Annahme , die Vorschrift **schütze** nicht die **Insolvenzmasse** , sondern nur die Gläubiger in ihrem Anspruch auf Gleichbehandlung in der Insolvenz .
  - 2 InsO) , so soll dieser Vorbehalt zwar die künftige **Insolvenzmasse schützen** , aber nicht zugleich das Vertrauen

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

[Ahmad et al. 1992] Khurshid Ahmad, Andrea Davies, Heather Fulford and Margaret Rogers (1992): “What is a term? The semi-automatic extraction of terms from text”, in: Mary Snell-Hornby et al.:*Translation Studies – an interdiscipline*, John Benjamins Publishing Company (Amsterdam/Philadelphia)

[Heid et al. 2008] Ulrich Heid, Fabienne Fritzinger, Susanne Hauptmann, Julia Weidenkaff and Marion Weller (2008): “Providing corpus data for a dictionary of German juridical phraseology”, in: Angelika Storrer, Alexander Geyken, Alexander Siebert and Kay-Michael W¨urzner: *Text Resources and Lexical Knowledge* (= Proceedings of the 9th Conference on Natural Language Processing, KONVENS 2008).

[Schiehlen 2003] Michael Schiehlen (2003): “A Cascaded Finite-State Parser for German”, in: Proceedings of the Research Note Sessions of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL’03), Budapest, April, 2003.