Legal Knowledge and Information Systems
F. Bex and S. Villata (Eds.)
IOS Press, 2016

doi:10.3233/978-1-61499-726-9-195

Computer Assisted Legal Linguistics (CAL²)

Hanjo HAMANN[a,1], Friedemann VOGEL[b] and Isabelle GAUER[b]
[a] *Max Planck Institute for Research on Collective Goods, Bonn, Germany*
[b] *Albert Ludwig University, Institute of Media Culture Science, Freiburg, Germany*

Abstract. We introduce Computer Assisted Legal Linguistics (CAL²) as a semi-automated method to "make sense" of legal discourse by systematically analyzing large collections of legal texts. Such digital corpora have been increasingly used in computational linguistics in recent years, as part of a quantitative research strategy designed to complement (rather than supplant) the more qualitative methods used hitherto. This use of statistical algorithms to analyze large bodies of text meets with an increasing demand by lawyers for empirical data and the recent turn towards evidence-based jurisprudence. Together, these research strands open exciting avenues for research and for developing useful IT tools to support legal decision-making, as we exemplify using our reference corpus of about 1 billion tokens from the language of German jurisprudence and legal academia.

Keywords. Computational linguistics, corpus linguistics, legal semantics, law and language, CAL²

1. Introduction

Law is performed in and through language, so both fields of study are intricately linked. They share various epistemological challenges [1] and have, incidentally, also undergone similar methodological changes in recent years: Both legal research and linguistics began to turn from largely introspective, intuitively-theorizing humanities into more empirical, evidence-based social sciences.

In linguistics, both the availability of computers for sophisticated statistical analyses and the surge of digital mass media as a resource for studying language phenomena have conspired to create a new discipline: Computational corpus linguistics. Being a quantitative approach to *social usage patterns* as the units underlying the evolution of natural languages, this discipline has made notable forays into the legal domain by considering law as a "sediment" of previous discourse patterns [2][3]. This enables legal linguists of the new quantitative variety to analyze legal phenomena, like other language phenomena, by algorithmically searching for and analyzing recurrent speech patterns in large machine readable collections (corpora) of legal text.

Similarly, legal researchers have turned to empirical methods and hard data for a more rigorous methodology when it comes to analyzing legal issues. The "New Legal Empiricism" [4] or, as it became known in Germany, "Evidence-Based Jurisprudence" [5], seeks to put legal arguments on a solid empirical footing by using empirical data, statistical analysis and meta-studies of social science research to improve legal decision-

[1] Corresponding Author: hamann@coll.mpg.de.

making. This movement has most recently given rise to a new kind of legal informatics that (unlike legal cybernetics in the past) seeks to answer epistemological questions of law with the assistance of computers (e.g., for Germany, www.en.lexalyze.de). This gives rise to what a recent paper in a law and technology journal aptly called "Big Data Legal Scholarship" [6].

Joining the forces of both these disciplines, and exploiting their similar recent developments, we introduce Computer Assisted Legal Linguistics (CAL²) as a novel approach to legal semantics. Being a crossroads of the quantitative empirical strands of both legal and linguistic research, CAL² opens up exciting new avenues of research, which are currently being explored by the International Research Group CAL² [7], based at Freiburg and Bonn (Germany) and funded by the Heidelberg Academy of Sciences (www.cal2.eu).

2. Project Progress

The CAL² Research Group has built a large "CAL² Corpus of German Law" (JuReko) which contains various types of texts from German jurisprudence and legal research, balanced in a so-called "reference corpus" which, like its counterparts in other language domains (e.g., the Corpus of Contemporary American English, COCA), allows for reliable big data analyses by corpus linguists, and for the development of new software tools to facilitate and improve legal decision-making.

The CAL² Corpus of German law contains all 6,300 German federal statutes (~2.3 M token) along with 370,000+ court decisions (~800 M token) and 43,000+ academic research papers (~150 M token). These texts were collected from various digital sources and prepared for computer-linguistic analysis by first extracting their metadata and storing them in a relational database of the following structure:

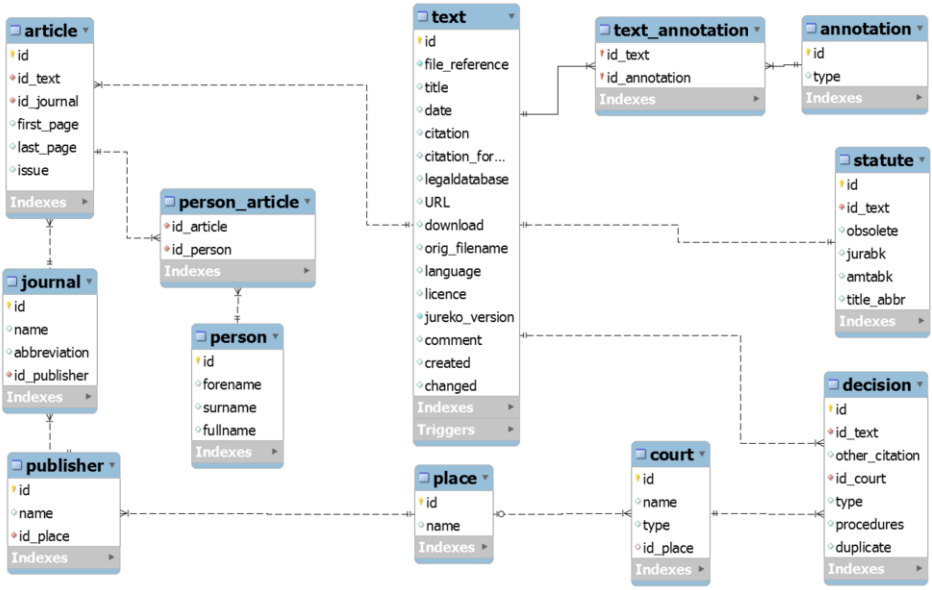

Figure 1. Relational database model of the "CAL² Corpus of German Law" (*JuReko*).

Simultaneously, our full texts were converted via jTidy to well-formed xml files. A pipeline of xsl transformations (tailored to the underlying source file templates) then converted the documents to TEI P5 compliant xml (www.tei-c.org/Guidelines/), being a de facto linguistic standard [8], particularly for text structures and metadata. Additionally we enriched our texts with part-of-speech information, annotated using TreeTagger [9].

To minimize errors during annotation, it was preceded by a labor-intensive vetting process conducted by hand. Specifically, metadata were checked for inconsistencies and normalized where possible. Where author information could not be extracted automatically from the source data, this was done manually.

Our automated text annotation was subsequently verified in several cycles of random tests. Various errors that had already plagued our data sources were successfully corrected along the way. Furthermore, duplicates were removed from the corpus.

The corpus, as a resource for subsequent computation and statistical analysis, can thus be relied on to be sufficiently well-kempt.

3. Pending Steps

As a next step towards analyzing the corpus and realizing its full potential, our research group will develop software tools (based on Java) to allow for data exploration and clustering-based analysis.

Specifically, based on a multilayer linguistic model, the software may generate co-text profiles for each of the 200,000 most frequent tokens and n-grams (where n = {2, 3, 4, 5}) which can be browsed in a Keyword-In-Context (KWIC) display, and submitted to significance testing (LLR, χ^2), contrasting occurrences both within-corpus and between our corpus and the general reference corpus for German language (DeReko) developed at the IDS Mannheim. This will allow us to measure quite precisely and under carefully controlled conditions how the usage of a certain token or n-gram varied in time and domain, subject area and text type.

Eventually, our software may be able to quantify and compare the degree to which the usage of a certain expression is fixed (as a “set” phrase) in the language of lawyers. We can thereby subject to an empirical test the notions of “rigidity” and “vagueness” that philosophers and linguists of law have developed in the past [10][11]. This strategy will fruitfully complement the qualitative approaches to legal language that have hitherto dominated the discourse.

As a further step into the future, our corpus might be fitted with a GUI to be used by other researchers or the general public. Owing to copyright restrictions, the corpus in its entirety cannot be released or licensed (as is true for its general language counterpart DeReko), but we are presently exploring options to create user interfaces or APIs. In the course of these plans, for which we currently solicit funding, we will also review other available representation standards, like CEN MetaLex (www.metalex.eu) and Akoma Ntoso (www.akomantoso.org) to ensure interoperability of our tools and compatibility of our corpus with other work from AI and law on the semantic web.

4. Conclusion

We propose a new approach to epistemological questions of language and law. By developing tools for computational assistance, we seek to address core philosophical

questions about the rigidity or vagueness of legal language. This empirical big data strategy, which complements (not supplants) traditional qualitative theorizing will expand our perspective on many problems in the legal arena. Providing user-friendly tools to explore and to statistically analyze the huge text corpora involved will mark the next step towards a future of Computer Assisted Legal Linguistics (CAL²).

References

[1] F. Vogel, H. Hamann, D. Stein, A. Abegg, L. Biel and L. M. Solan, *"Begin at the beginning". Lawyers and Linguists Together in Wonderland*, Winnower **3** (2016), no. 4919.
[2] F. Vogel, *Linguistik rechtlicher Normgenese. Theorie der Rechtsnormdiskursivität am Beispiel der Online-Durchsuchung*. De Gruyter, Berlin, 2012.
[3] F. Vogel, Das Recht im Text. Rechtssprachlicher Usus in korpuslinguistischer Perspektive. In: E. Felder, M. Müller and F. Vogel (eds.): Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analysen, De Gruyter, Berlin, 2012, 314–353.
[4] M. C. Suchman, and E. Mertz, Toward a New Legal Empiricism: Empirical Legal Studies and New Legal Realism, *Annual Review of Law and Social Science* **6** (2010), 555–579.
[5] H. Hamann, *Evidenzbasierte Jurisprudenz. Methoden empirischer Forschung und ihr Erkenntniswert für das Recht am Beispiel des Gesellschaftsrechts*, Mohr Siebeck, Tübingen, 2014.
[6] F. Fagan, Big Data Legal Scholarship: Toward a Research Program and Practitioner's Guide, *Virginia Journal of Law & Technology* **20** (2016), 1–81.
[7] C. Coupette, Legal Tech Will Fundamentally Change Legal Research – Interview With Dr. Hanjo Hamann, *Legal Tech Blog* (http://legal-tech-blog.de/legal-tech-will-change-legal-research), 16 Feb 2016.
[8] M. Stührenberg, The TEI and Current Standards for Structuring Linguistic Data: An Overview, *Journal of the Text Encoding Initiative* **3** (2012), online at DOI 10.4000/jtei.523.
[9] H. Schmid, Improvements in Part-of-Speech Tagging with an Application to German, *Proceedings of the ACL SIGDAT-Workshop*, Dublin, 1995.
[10] G. C. Christie, *Vagueness and Legal Language*, Minnesota Law Review **48** (1964), 885–911.
[11] G. Keil and R. Poscher, *Vagueness and Law. Philosophical and Legal Perspectives*, Oxford University Press, in print 2016.