# Computer Assisted Legal Linguistics (CAL²): An interdisciplinary approach

Yinchun Bai (University of Freiburg, Germany), Isabelle Gauer (University of Freiburg, Germany), Hanjo Hamann (MPI Bonn, Germany) and Friedemann Vogel (University of Freiburg, Germany)

## Introduction

In law as well as in linguistics researchers work with texts. In recent years both disciplines are turning away from the introspective approach to a more empirical, evidence-based practice. Digital sources and programmes for statistical analyses of mass data have become increasingly available and offer new opportunities for linguistic research. In law and legal linguistics a growing number of researchers and practitioners are also acknowledging the merits of working with large text collections, so they promote steps towards an evidence-based jurisprudence (Goźdź-Roszkowski, 2011; Mouritsen, 2011; Hamann, 2014; Fagan, 2016; Solan, 2016; Vogel, 2017). "Computer Assisted Legal Linguistics" (CAL$^2$) (Vogel, Hamann, and Gauer, in press) tries to support the process of legal problem-solving by using corpus linguistic methods. That means, computer-supported analysis of carefully pre-processed corpora of legal texts are used to analyse legal semantics, language and sociosemiotics in different working contexts (judiciary, legislature, legal academia, etc.). It does not seek to replace hermeneutic procedures with algorithms but to complement them. Empirical data and computer algorithms can only support legal work, but decision making will always be cognitive processes of contextualization.

## The CAL² corpus of European law

The CAL²-group seeks to bring together members from the fields of law, linguistics and computer science and to build an infrastructure for legal linguistic analysis. To do this, the first step was to create a corpus that is suitable and balanced enough to address questions on the language and law interface. That means, it must contain texts that exhibit the special properties of the legal genre. At the moment, we have included German court decisions, statutes and articles from legal journals and English court decisions as the starting point of building a corpus of European Law. To be more specific, the CAL² corpus of European law now contains: 6,300 German federal statutes (~15 M words), 370,000 court decisions in German (~800 M words), 20,000 court decisions in English (~90 M words) and 43,000 German academic research papers (~200 M words), which equal over 1 billion words in total (see Figure 1 below for the temporal distribution of the data).

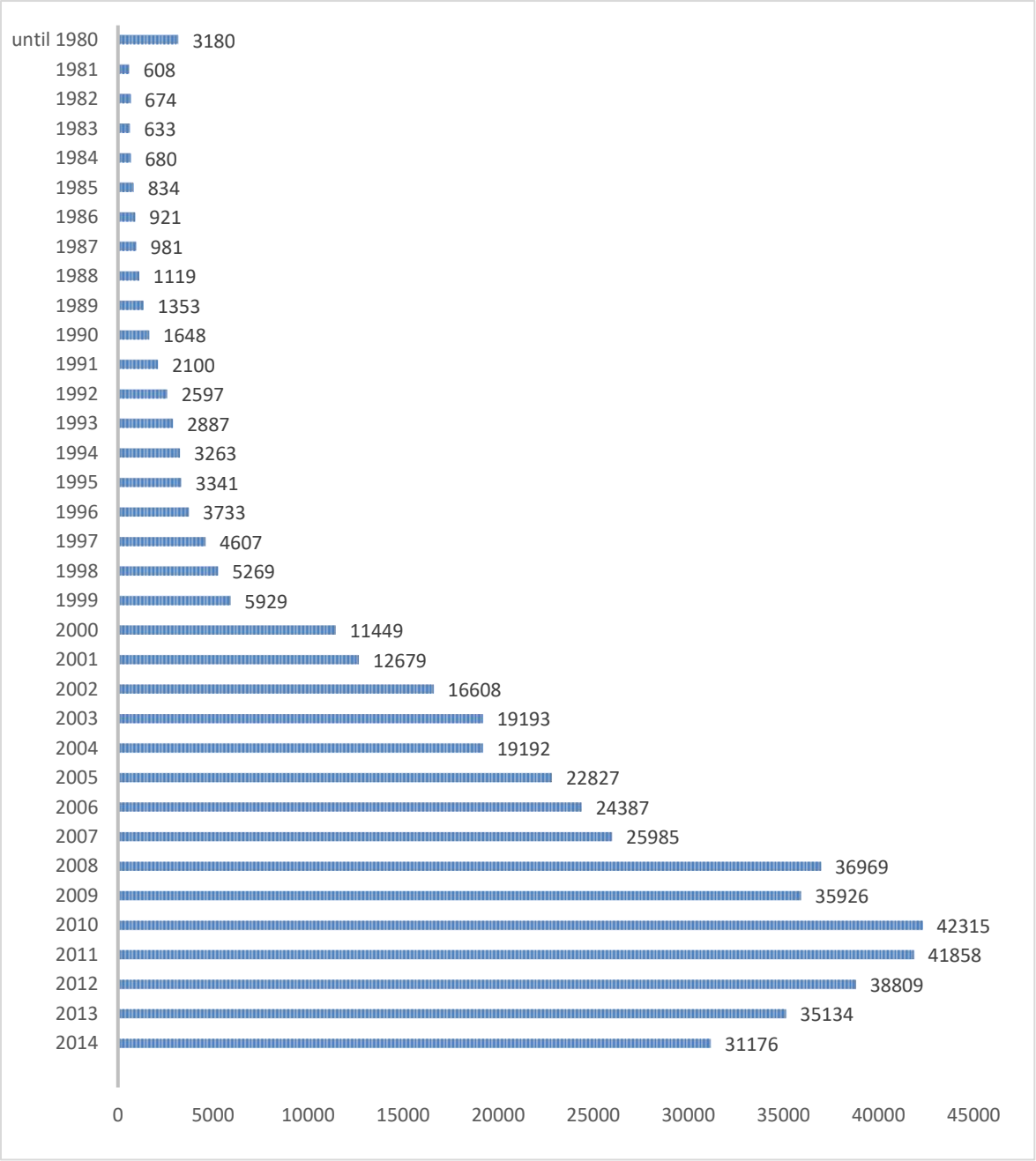


*Figure 1 number of texts per year*

The collected texts were converted to TEI P5 (www.tei-c.org/Guidelines/) compliant xml which serves as a de facto standard for text structural annotation (Stührenberg, 2012). A pipeline of xsl transformations tailored to the different types of input formats was created to construct the corpus. Additionally, we enriched the data with part-of-speech annotation using TreeTagger (Schmid, 1995). During the corpus building, error control and duplicate removal were integrated in the process. It was important to get the data as clean as possible to be able to conduct quantitative as well as qualitative analysis. To assist the cleansing work and to get a better overview of the corpus content, the metadata of the legal texts were also collected and documented in a related database.

## Pending steps

The next step is the development of a platform for specialised statistical processing that supports legal linguistic research and further also legislation and decision making (for example statutory interpretation of ambiguous expressions). The platform will help to gain insight to questions like how legal terms were used in the past and today, how often and by whom. This can assist the research on diachronic changes of linguistic expressions in law and the development of dogmatic schools.

The implementation of the platform will comprise three parts: First, frequency lists will be created to derive the 200,000 most frequent lemmas (nouns/verbs/adjectives) of the corpus. For all of them we will calculate context profiles for statistical multi-level-context-analysis that contain co-occurrences and usage patterns according to the metadata (for example when the expression was used most frequently). Secondly, we attempt to measure the rigidity of expressions to calculate how fixed or varied the expressions are in different contexts. Lastly, we examine semantic similarity in legal texts by clustering similar context profiles and visualizing the results as self-organizing maps.

## Case study

To test the CAL$^2$ corpus in action and to explore its applicational possibilities and limitations, we conducted a case study using parts of the corpus data. The case study is a comparative study of the linguistic formulation of the "*employee*" concept in UK and EU court decisions. The choice of the term "*employee*" as our study object was motivated by several considerations. First, the employment law is one of the most dynamic legal domains. Since it does not regulate the behaviour of the legal subjects directly, but confines and develops individual rules which in turn regulate the relevant affairs, the working techniques involved are complex and hardly have legal benchmarks to compare with (Vogel, Pötters, and Christensen, 2015). The development of terminology in this legal domain is therefore especially complicated, which made the study of an employment law-related notion highly interesting. Second, the term "*employee*" is an especially important concept in the framework of the employment law, because the employment law is by nature the employee protection law (Hueck and Nipperdey, 1957) which mediates the relationship between employees, employers, trade unions and the government.

## Data

Our research material of this study comes from the English part of the CAL$^2$ corpus. It consists of judgements made by the UK Employment Appeal Tribunal [UKEAT] and the employment law-related judgements from the Court of Justice of the European Union [EUECJ]. Using texts of judgement to study the use of a certain term underlies the representativeness of our findings. On one hand, the terms used in judgements are linguistically produced after conceptualization and interpretation by judges, who represent the understanding and use of a term both as a regular native speaker and as a legal professional – they are aware of the definition and the legal force of a term under the legal context but are still confined by the limitations of individual intuition about the frequency and the pragmatic implications of the usage. This helps to reflect the conception of "*employee*" at the dynamic user level. On the other hand, court

decisions are a specialized type of text production, which are highly contextualized and constrained by the institution of law. The formulation of language involving the use of "*employee*" is thus subject to specially institutionalized format as well as semantics according to the corresponding legal conventions. They therefore help to reveal the conception of "*employee*" at the underlying institutional level.

## Research questions

This study addresses the following research questions: (1) What are the central "*employee*" concepts and "*employee*"-related concepts within the labour law framework in the UK case law system? (2) How are the "*employee*" concepts addressed in UK court decisions in terms of frequent usage patterns? (3) How similarly and differently are the "*employee*" concepts linguistically formulated and presented in EU court decisions?

## Results

In terms of the frequency distribution of various "*employee*" concepts and the usage patterns of the term, we found both commonalities and differences in UK and EU court decisions. This is achieved through analysing the compounding and phrasal structures, the genre-specific co-occurrence partners, and the predication patterns of the term "*employee*". In both UK and EU court decisions the majority of compound nouns of "*employee*" concern the status of a person in an employment relationship, i.e. if a person can be classified as an employee at all (e.g. *ex-employee*), although the EU court decisions exhibit a much lower rate of using compound nouns in general. For noun phrases containing "*employee*", we identified four phrasal structures, including modifier-noun phrases, noun-noun phrases, possessive-noun phrases, and *of*-possessive phrases. Generally speaking, the modifier-noun phrases either identify the types of an employee in relation to the employment relationship or describe aspects of the employee proper, such as "*full-time employee*" and "*female employee*"; the noun-noun phrases, possessive-noun phrases, and of-possessive phrases, on the other hand, mostly address the expansive aspects of an employee, including e.g. people related to an employee ("*employee representative*"), and external components and events involved in an employment relationship ("*employee's contract*", "*dismissal of employee*"). From the semantic perspective, the "*employee*" concept in both UK and EU legal cultures has a core semantic composition of "temporal specification of an employment relationship", "discrimination between social groups (gender, age, etc.)", and "rights and entitlement", as reflected by the frequent co-occurring partners in phrasal structures, expanded contexts [+/-15], and predication structures. But the UK and EU court decisions do show different emphases on a few specific aspects of the employee and the employment relationship. For example, the UK court decisions contain more predications describing an employee's conducts and conditions in an employment relationship, while the EU court decisions feature a slightly bigger percentage of notions that concern the support of an employee.

## Conclusion

To summarise, the CAL$^2$ research group engages itself in corpus-assisted legal linguistics and strives to contribute to this young and vibrant working field by compiling

a comprehensive and balanced specialised corpus of legal texts and developing accompanying analytical tools. Our approach aims at making the possible interpretations more visible and thus making the law more transparent by algorithmically searching for and analysing recurrent speech patterns in large linguistic corpora of legal texts. With a case study analysing the conception and usage patterns of the legal term "*employee*", we showed one of the practical uses of the CAL$^2$ corpus and experimented on the possibility of identifying conceptual sediments that exist as constants in the European way of legal thinking.

## References


Fagan, F. (2016). Big data legal scholarship: Toward a research program and practitioner's guide. *Virginia Journal of Law & Technology*, 20, 1–81.

Goźdź-Roszkowski, S. (2011). *Patterns of linguistic variation in American legal English: A corpus-based study*. Frankfurt am Main, New York: Peter Lang.

Hamann, H. (2014). Unpacking the Board. A Comparative and Empirical Perspective on Groups in Corporate Decision-Making. *Berkeley Business Law Journal*, 11, 1–54.

Hueck, A., & Nipperdey, H. C. (1957). *Lehrbuch des Arbeitsrechts*. Berlin: Vahlen.

Mouritsen, S. C. (2011). Hard cases and hard data: Assessing corpus linguistics as an empirical path to plain meaning. *Columbia Science and Technology Law Review*, 33, 156–205.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to german. *Proceedings of the ACL SIGDAT-Workshop*. Dublin.

Solan, L. M. (2016). Can corpus linguistics help make originalism scientific. *Yale Law Journal Forum*, 126, 57–64.

Stührenberg, M. (2012). The TEI and current standards for structuring linguistic data: An overview. *Journal of the Text Encoding Initiative*, 3. Retrieved from http://jtei.revues.org/523

Vogel, F. (2017). Calculating legal meanings? Drawbacks and opportunities of corpus assisted legal linguistics to make the law (more) explicit. In D. Stein & J. Giltrow (Eds.), *The pragmatic turn in law. Inference and Interpretation*. New York, Boston: Mouton de Gruyter.

Vogel, F., Hamann, H., & Gauer, I. (in press). Computer assisted legal linguistics: Corpora and empirical methods as a new instrument in the legal toolbox. *Law & Social Inquiry. Journal of the American Bar Foundation (ABF)*.

Vogel, F., Pötters, S., & Christensen, R. (2015). *Richterrecht der Arbeit - empirisch untersucht. Möglichkeiten und Grenzen computergestützter Textanalyse am Beispiel des Arbeitnehmerbegriffs*. Berlin: Duncker & Humblot.