Calculating legal meanings?

Drawbacks and opportunities of corpus assisted legal linguistics to make the law (more) explicit

Friedemann Vogel

1. Introduction

In the following contribution I would like to introduce the potentials and the challenges of corpus assisted legal linguistics. By that I mean computer supported qualitative discourse analysis of legal texts and legal semantics. For this purpose I first give a short introduction to my working area of legal linguistics (especially grounded in the discussions of Germany) and my understanding of jurisprudence as a text based institution (2). Then secondly, I show how corpus linguistics can help us to see legal discourses in a new perspective, and give us methods to analyse speech patterns as indicators of sediments of legal dogmatic (3). Thirdly, I demonstrate the approach of legal corpus pragmatics using the example of the expression *employee* (*Arbeitnehmer*) in a corpus of labour court decisions (4). In my conclusion (5) I try to summarize the drawbacks and opportunities of these methods and present a new research project to develop a German legal reference corpus.

2. Law as institutionalized textual work

The given topic of interest is located in an interdisciplinary working group in the south of Germany, called *Heidelberg Group of legal linguistics* (Heidelberger Gruppe der Rechtslinguistik). This research group was founded in 1984 by the lawyer, Professor Friedrich Müller and the Linguist, Professor Rainer Wimmer. They both had formed an interest in the relationship of legal system, language and speechacts long before they finally met. But they came from different fields, with different theoretically backgrounds and methods. Together with other colleagues they have developed a common theory of legal linguistics over the last thirty years. Today the members of the group come from very different backgrounds, legal scholars and linguists, judges and language practitioners, philosophers and former presidents of parliament. The common understanding of most of these legal linguists – like me – might be described in the following three principles:

a) In modern societies, legal work is work within texts and language. When lawyers work with norms they actually work with many different texts. They connect a text with others texts, for example statutes with prior court decisions, texts of the legal scientific community, legal commentaries, texts of external opinions of experts, and, of course, texts describing the controversial "real facts". In other words: The modern constitutional state establishes an intertextual structure (Müller, Christensen, and Sokolowski 1997, Morlok 2004). This is not just another attribute among others. The constitutional state *is* indeed a text structure in itself. Jurisprudence is a "text-based decision science" ("textbasierte Entscheidungswissenschaft", Morlok in press). This relationship of legal system and text depends on two functions: Language is the most important medium to share and negotiate legal norms as behavioural expectations under a threat of penalty. Furthermore and more important in my opinion, language based constitutional democracy transforms the brute force of social conflicts into due process and a semantic struggle to find better arguments.

b) We have to distinguish between at least three levels describing language use: (i) The *surface of language* as sensual stimulus hints, that are expressions like words, phrases, texts, intertextual structures, but also facial expressions and gestures, clothes and architecture forms. In terms of de Saussure but in a broader sense this level can be called "signifiant". On the level of the "signifiant" a

sign is constituted if we can associate something with meaning or (ii) *cognitive concepts*. In the words of Lawrence Barsalou: „By concept I mean the descriptive information that people represent cognitively for a category, including definitional information, prototypical information, functionally important information, and probably other types of information as well. In this regard, my use of concept vaguely resembles intension and sense." (Barsalou 1992: 31)

These concepts or frames (Goffman 1974; Minsky 1975) or, in terms of de Saussure, the "signifié", are dynamic and have been built "bottom up" on base of sensory perception and socialization. Concepts also preform our perception top down. When the top down process predominates, then we speak of stereotypes or – in a negative form – common prejudice (Nelson 2006). Beside the surface of language (or expressions) and cognitive concepts (or knowledge frames) we must distinguish (iii) the *social practises* constituting the connection between both of them, that is processes of contextualization in the sense of John Gumperz (1982) or Peter Auer (1986). John Gumperz called it a "paradox": „To decide on an interpretation, participants must first make a preliminary interpretation. That is, they listen to speech, form a hypothesis about what routine is being enacted, and then rely on social background knowledge and on co-occurrence expectations to evaluate what is intended and what attitudes are conveyed." (Gumperz 1982: 171)

Gumperz points out that interpretation of an expression depends on our individual experience with its use in the past, our knowledge about situations and the people involved. In that sense we can say, we make expressions or "contextualization cues" (Gumperz 1982: 131) "meaningful" (Hörmann 1980) within a virtual framework or cognitive context model (van Dijk 1999). But we have to consider, that contextualization and the production of meaning is not only a question of individual cognition, but of *social* cognition (Schützeichel 2007). Meaning is neither external nor only internal; it is a virtual product of interaction between human beings (Goffman [1967] 2002). In this perspective we have to focus on the procedures of how people learn and constitute meanings in different situations. Or in short: interaction between family members differ from the interaction in institutions like the jurisprudence. That's why we speak about legal language as a technical terminology. The main task for students of law is learning an adequate use of this terminology, how to connect and write legal texts with institutionalized methods and how to behave in accepted rituals.

c) Consequently, as the third principle, we must distinguish the "text form" of a norm (for example laws, decisions, and commentaries) and the norm as a concept. Norms cannot be inside of a text, ready for use and ready for any subsumption. The law is not a "pot" (Busse 1992: 14). Lawyers must construct a norm actively in a hermeneutic way accepted in the community of lawyers that is with arguments for their interpretation. They have to contextualize and concretize the expressions of a clause with the help of other texts; they must *ascribe* a norm to a text of law. You can observe these processes especially on the other side of norm text creation. In fact, legal colleagues involved in the legislative process do not try to create one meaning of a law. Instead they try to anticipate the main addressees, their previous knowledge of language and norm structure. They try to anticipate potential different contextualization of the arising text (Vogel 2012a). As a result, we can observe three analytical levels: the world of norms, the world of things (or *Lebenswelt*) and the world of texts. Both the world of norms and the world of things must be constituted by the world of texts. Jurisprudence, in my opinion, is the most important institution of our society to develop methods to bring these three worlds together and in a constitutional (hierarchical) order. In my view, this is also the main question and challenge for legal linguistics: How can we make legal interpretation more

transparent? How can we make implicit text connection and legal inferences more explicit, especially in a digitalized world where thousands of texts are available in databases?

These last questions suggest legal interpretation is not being transparent. And often, when legal linguists in Europe talk about discursive construction of legal norms and of normativity within text, some legal scholars disagree. In their view the word "construction" is associated with despotism. But in fact, jurisprudence has developed and used a complex set of institutionalized interpretation methods. For example, Savigny's Canons – grammatical, historical, systematical and theological interpretation – are arguments for expanding and reducing legal meanings; they place an expression in its particular speech context (Kudlich and Christensen 2004: 83). Also theories to structure a hierarchy of legal texts try to bring interpretation under professional control.

However, in that long tradition of methodology of legal text interpretation the important role of introspection often seems to be forgotten. Introspection means, that the hermeneutic process depends only on the individual speech experience of the interpreter. In everyday life introspection is important for the ability to judge quickly and effectively. But there are also problems described by cognitive psychology under the topic biases of heuristic in judgement and decision-making (cf. Starck and Deutsch 2002). Especially if we try to make intuitive statements about the frequency of something, for example the use and the meaning of a word in different contexts, we often fail. We can test it, for example, with the following questions:

- How many words can native speakers produce actively and how many words can we understand passively?[1]
- What is the most frequent used word in our or in your language?[2] And which frequency order would you give for the words *candle-light, Gregorian* and *police car*[3]?
- And how many and which meaning versions of the expressions *ship* and *old* do you know?[4]

In short: Introspection is a necessary, but not a sufficient resource for certain interpretations. This fact is the starting point of corpus linguistics.

3. Corpus linguistics and legal corpus pragmatics

To control introspection corpus linguists usually use selected text data bases and combine hermeneutic methods with computer assisted analysis methods. By now there are huge text collections (text corpora) for analysing standard languages. Such reference corpora, trying to capture

[1] It is very difficult to calculate frequencies of entities in the mental lexicon, we can only estimate them. The biggest problem is to separate so called common language and different institutionalized terminologies. For example, for adults we accept 300-500 thousand words of common language of German. But in everyday life we speak only 12-16 thousand words and understand at least 50 thousand words (as a native speaker); cf. Best 2000, 2006.

[2] In German and English the most frequent words are articles (*der/die/das* or *the*).

[3] In German (*Kerzenlicht, Gregorianisch* and *Polizeiwagen*), all these words are in the same frequency-class 17, that is, these words belong to the least frequent expressions in contrast to the most frequent word (articles) in the German Reference Corpus (cf. Korpusbasierte Wortgrundformenliste DeReWo, v-30000g-2007-12-31-0.1, mit Benutzerdokumentation, http://www.ids-mannheim.de/kl/derewo/, © Institut für Deutsche Sprache, Programmbereich Korpuslinguistik, Mannheim, Deutschland, 2007).

[4] The corpus driven made online dictionary *Elexico* of the Institut für Deutsche Sprache (Mannheim, Germany) counts five different meanings of the German word *Schiff* (*ship*): ›Wasserfahrzeug‹ (water craft), ›Kircheninnenraum‹ (nave), ›Wasserbehälter‹ (water reservoir), ›Teil des Webstuhls‹ (shuttle of a weaving loom) und ›Zinkplatte‹ (slide of zinc) (http://www.owid.de/artikel/87462, 19.05.2013). For the expression *alt* (old) Elexico counts seven different meanings (http://www.owid.de/artikel/271695, 19.05.2013).

almost all relevant contexts of a language, exist for different languages including German, French, British and American English. For example the British National Corpus (BNC) consists of a 100 Million words, the German Reference Corpus (DeReko) even takes about 8 Billion words. However, these corpora are not useful to analyse questions about legal language. And specialized legal text corpora do not exist yet. At the moment I am preparing a Legal Reference Corpus, but we will come to that later. What can we do with these corpora?

Corpus linguists develop and use special algorithms and tools realizing one great purpose: They automatically collect those statistically relevant speech patterns, which will be overseen by purely introspective analysis. I will give a brief introduction to the most important algorithms:

– Most systems allow complex search tasks, using special search syntax as *regular expressions*. The results are presented in the form of *concordances* that is the "search term" or *keyword in context* (KWIC), as it is realized in the freeware tool *AntConc* (Anthony 2005): Here you can see the search term – for example the expression *arbeitnehm* (employee) in the middle of the window and the particular context right and left. If you sort the results alphabetically you can see recurrent speech patterns distributed in the corpus.

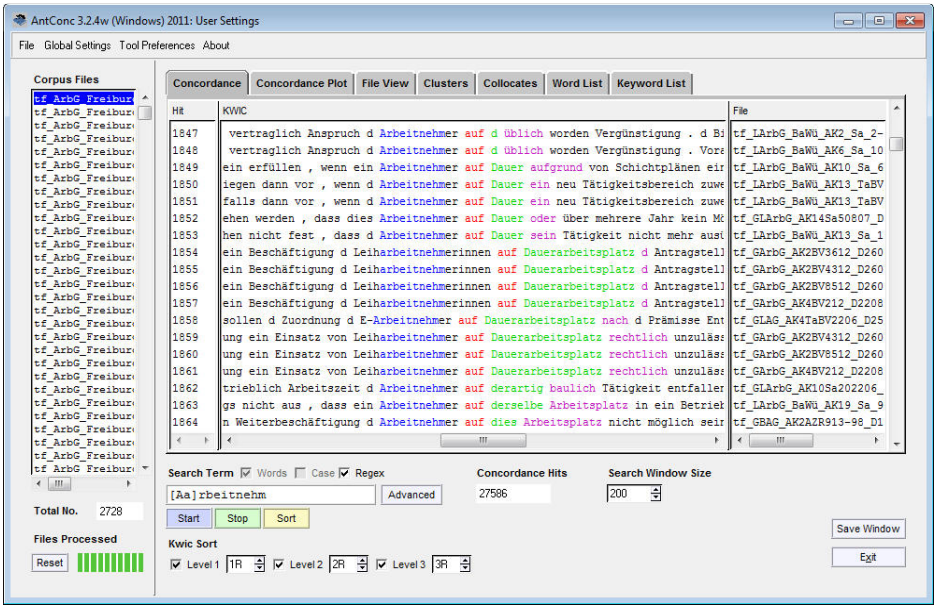


Fig. 1: Concordances in the freeware tool AntConc (Anthony: 2005)

– Furthermore you can create *word lists*, collecting all expressions given in a corpus and count them. Word lists taken separately are hardly meaningful. More interesting is the statistical comparison of two different word lists to get specific *keywords*. For example you can compare the word list of a legal text corpus with the word list of a corpus of newspaper texts. As a result you will get these words which are typical for the one corpus but atypical for the other. For example, see below two keyword lists of German texts (using the toolkit for linguistic discourse and image analysis, LDA-Toolkit, cf. Vogel 2012c), contrasting expressions used from supporter and opponents of nuclear power. As a result we see expressions typically used in arguments as buzzwords or shibboleth.

Table 1: Comparing keywords in the LDA-toolkit (Vogel: 2012c)

expression	translation	X^2	conservative parties		left-winged parties	
			f	f/10.000	f	f/10.000
FDP	[German liberal party]	168,8	84	146,9	22	12,7
Kernkraftwerk	nuclear power plant	79,5	37	64,7	8	4,6
CDU	[German conservative party]	77,6	34	59,4	6	3,4
ausreichend	sufficient	60,3	28	48,9	6	3,4
Übergangstechnologie	transition technology	52,7	19	33,2	1	0,5
zur	for	52,4	40	69,9	22	12,7
Kernenergie	nuclear energy	51,3	30	52,4	11	6,3
Energiemix	energy mix	43,7	16	27,9	1	0,5
bezahlbar	affordable	37,7	14	24,4	1	0,5
Brückentechnologie	bridging technology	37,7	14	24,4	1	0,5
als	when/as	36,5	44	76,9	38	21,9
BüSo	[small German party]	33,2	11	199,2	0	0
klimafreundlich	climate-friendly	33,2	11	19,2	0	0
Laufzeit	run term	32,5	20	34,9	8	4,6
aber	but	31,2	34	59,4	27	15,6
Neubau	new building	29,8	18	31,4	7	4
können	could	29,2	55	96,2	64	37
lehnen	lean	27,2	13	22,7	3	1,7
Teil	part	25,7	10	17,4	1	0,5
verfügbar	available	24,8	11	19,2	2	1,1
Grüne	[German green party]	48,9	0	0	147	84,9
Linke	[German left-winged party]	35,9	1	1,7	115	66,4
Risiko	risk	29	2	3,4	101	58,3
Atomkraft	nuclear power	25,9	18	31,4	178	102,9
SPD	[German labour party]	23,5	0	0	71	41
ungelöst	unsolved	23,5	0	0	71	41
zudem	furthermore	20,2	0	0	61	35,2
Endlagerfrage	problem of final storage	19,8	0	0	60	34,6
vereinbart	agreed	16,8	0	0	51	29,4
früher	earlier	16,8	0	0	51	29,4
keinesfalls	on no account	16,5	0	0	50	28,9
stoppen	stop	15,8	0	0	48	27,7
deshalb	therefore	15,5	2	3,4	60	34,6
besonders	especially	14,9	1	1,7	52	30
Restlaufzeit	remaining term	14,5	2	3,4	57	32,9
Atommüll	nuclear waste	13,6	2	3,4	54	31,2
Atomausstieg	denuclearization	13,4	1	1,7	47	27,1
unverantwortlich	irresponsible	13,2	0	0	40	23,1
wollen	want	13,2	6	10,4	75	43,3
unsicher	unsafe	12,1	4	6,9	61	35,2

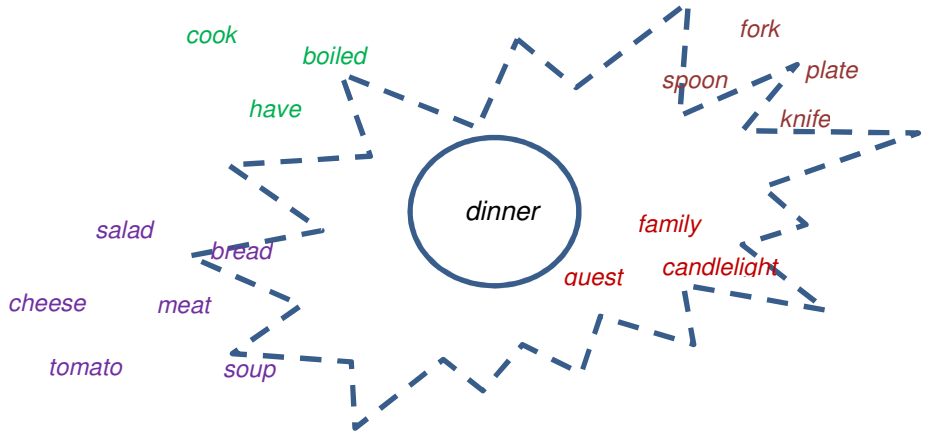
- You can also count *multi word units* (MWU) or n-grams to analyse stable phrases. For example, the algorithm separates all texts of a corpus in 5-grams (five-word-units) and creates a list containing the different expression types and their frequency. If you compare different lists, you can analyse *key-multi-word-units*, which are specific for the one and unspecific for the other corpus. See below, for example, a list contrasting three-word-units of left-winged parties and conservative parties in the same corpus of texts discussing pro and contra of nuclear power in Germany:

Table 2: Key-multi-word-units

expression	translation	X^2	conservative parties		left-winged parties	
			f	f/10.000	f	f/10.000
Ausbau der erneuerbaren	extensions of renewable	19,214	0	0	58	33,536
früher vom Netz	sooner off the net	16,558	0	0	50	28,91
die besonders unsicher	which are very unsafe	15,894	0	0	48	27,754
Atomkraftwerk noch früher	nuclear power plant even sooner	15,563	0	0	47	27,175
und die Endlagerfrage	and the issue of permanent disposal	15,563	0	0	47	27,175
besonders unsichere Atomkraftwerke	very unsafe nuclear power plants	15,563	0	0	47	27,175
das Risiko der	the risk of	14,988	1	1,75	52	30,066
die Endlagerfrage ungelöst	the issue of permanent disposal unsolved	14,899	0	0	45	26,019
Risiko der Atomkraft	risk of nuclear power	14,328	1	1,75	50	28,91
Atomkraft ist unverantwortbar	nuclear power can't be taken responsibility for	8,932	0	0	27	15,611
Atomkraft ist unverantwortlich	nuclear power is irresponsible	6,946	0	0	21	12,142
Endlagerung von Atommüll	permant disposal of atomic waste	2,975	0	0	9	5,204
Frage der Endlagerung	issue of permanent disposal	2,644	0	0	8	4,626
in erneuerbare Energien	in renewable energies	2,314	0	0	7	4,047
unser Kind und	our child and	1,983	0	0	6	3,469

– In my view the most important algorithm is the analysis of co-occurrences. Co-occurrences are words, which can be found in a defined context of a search expression more often than statistically expected. In simple terms: Co-occurrence algorithms look for the search term in a corpus and count, for example, all the words given on eight places left and right of the search term. The statistical information gives us the degree of cohesiveness between the words. In a pragmatic perspective co-occurrence analysis realizes in a technical and systematical way, what Ludwig Wittgenstein called "the meaning of a word is its use in the language" (PI 43). Co-occurrence profiles can be seen as context profiles of an expression. Finally, if you combine multi co-occurrence studies, you can explore the system of language use empirically step by step.

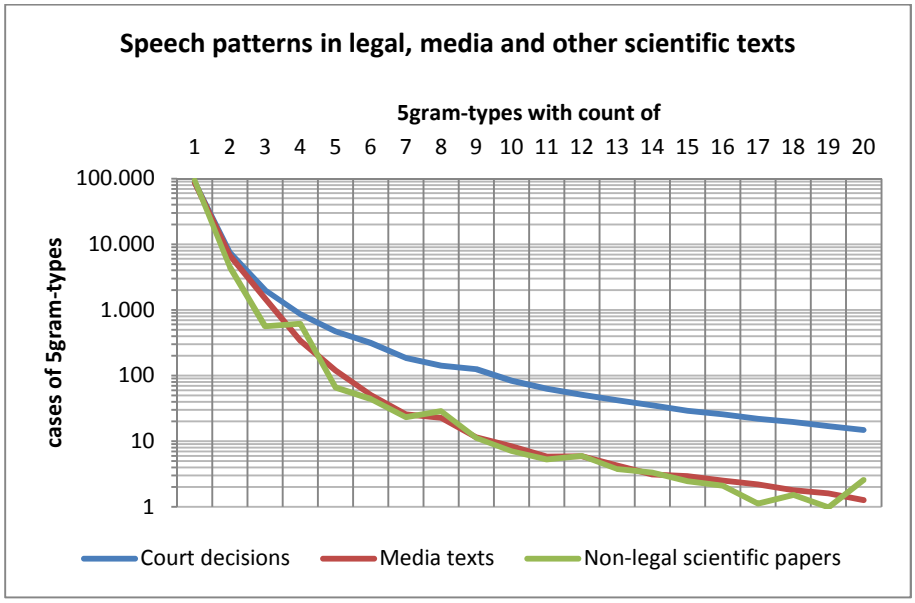
Fig. 2: Illustration of different co-occurrence instances exploring the context of the expression *dinner*



These algorithms or tools allow us to structure big text data and to form corpus assisted hypothesis about language use. I will use them in a framework of legal corpus pragmatics (Felder, Müller, and Vogel (Ed.) 2012; Vogel 2012b; Vogel (Ed.) in press) analysing legal speech patterns as a trace or as a trail of sediments of legal semantics. That assumption includes, that stable concepts of jurisprudence (dogmatic) are connected with a recurrent use of language. Furthermore, I would argue, that in the

current environment of an almost exponential growth of legal texts, recurrent speech patterns play an important role for a jurists' take on the state of the art. This means, speech patterns in legal contexts symbolise the dogmatic status or and the degree of institutionalisation of the concept associated with the expression through its "patterness" (Musterhaftigkeit). If the assumptions are correct we should see more patterns in legal texts than in media texts or other scientific texts. We can also test this question with corpus linguistic methods:

Fig. 3: Speech patterns in legal discourses as indicators of legal dogmatic



The chart above illustrates the degree of 'patterness' contrasting German court decisions, media texts and non-legal scientific texts. The compared corpora are:

- Corpus with 9.085 texts (court decisions) of German Labour Law (about 22,2 million words, 1950-2012; for more details see section 3.)
- Corpus with 474 non-legal scientific texts (papers) of Linguistics, Medical Science and Applied Ethics (about 2,6 million words, 1970-2012)
- Corpus with 17.974 texts of German media (about 10,9 million words, 2000-2012)

For that purpose I divided all texts into five-word-units and counted the hits of different unit types. As expected, using a particular five-word-unit (for example *this is a word unit*) exactly one time is the most frequent case for all three domains. But the case using a particular five-word-unit exactly in ten hits is found ten times more often in legal texts (about 100) than in media or non-legal scientific papers (about 10). That means that judges more often use recurrent multi-word-units in contrast to both of the other domains.

What do these patterns mean? If we look into the corpus and into the texts we can see that most of the speech patterns are references and citations of decisions of higher instances, especially governing arguments of the German Supreme Court (Bundesverfassungsgericht) or other federal courts. With a closer look, we can also see that different types of speech patterns constitute and are associated with different case groups. Of course, the patterns do not decide the case, but they give a macro structure for defining the topic more clearly.

In the following I will illustrate the methodological approach of Legal Corpus Pragmatics using the example of the word *Arbeitnehmer* (employee) in a corpus of labour court decisions[5]. The example is part of a pilot study of Ralph Christensen, Stephan Pötters – both lawyers – and me (in press). We wanted to test corpus linguistic methods in contrast and in addition to traditional legal methods. In this respect the meaning of *Arbeitnehmer* is a good test object, because the expression is not an expression of law but especially of case-law and dogmatic (so called "vague legal concept").

4. The concept of *Arbeitnehmer* (employee) in German labour law

To analyze the concept of *Arbeitnehmer* in labour law we built a corpus of 9.085 decision texts (22,22 Mio. words) of federal, county and European courts, published between 1954 and 2012.The texts had to be transformed in an simple-text format and annotated with part of speech information. For diachronic analysis the texts were separated in three groups:

(A) 1954 – 1989: 1320 texts / 1, 0 Mio. Words
(B) 1990 – 1999: 5036 texts / 13, 4 Mio. Words
(C) 2000 – 2012: 2728 texts / 7, 9 Mio. Words[6]

To explore the meaning of the expression *Arbeitnehmer* in the sense of Wittgenstein we used corpus linguistic methods to find speech patterns, framing the expression *arbeitnehm* in our corpus. We supposed that the expression *arbeitnehm* at least would be connected with the relevant concept, so called 'minimal assumption'. In the first step we collected several levels of context expressions realizing conceptual attributions of the type >X is Y<:

- Compounds like *Leiharbeitnehmer*, *Arbeitnehmerschutz*, *Arbeitnehmerüberlassung*, *Vollzeitarbeitnehmer*, *Fremdarbeitnehmer* etc.: the attribution (Y) is located in the determinans or determinandum;
- Predication phrases like **arbeitnehm** *ist* Y or *Arbeitnehmer ist nicht/kein* Y (employee is [not] Y);
- Attribution phrases with Adjectives like *junge/alte/gewerbliche/männliche* **arbeitnehm** (young/old/commercial/male etc. employee);
- Verb phrases with *arbeitnehm* like **arbeitnehm* müssen/können/dürfen* Y (employee must/are allowed to Y);
- Multi-Word-Units with *arbeitnehm*, for example *Arbeitnehmer im Sinne des* Y (employee in the sense of Y);
- Iterations of co-occurrences to the expression *arbeitnehm* (see below).

In the second qualitative step we grouped these context expressions into clusters of meanings or topics, grounded in qualitative micro analysis of the particular expressions in the texts. In the last step we built hypotheses about the most common framing structures associated with the expression *Arbeitnehmer* in contrast to other findings. Following I give some examples of the different categories:

(a) We found lots of **compounds** with the expression *arbeitnehm* and can distinguish between determinans and determinandum. The determinandum or primary word gives a basic frame; the

[5] For a first test of legal corpus pragmatics using the example of the expression *human dignity* (Art. 1 Abs. 1 Grundgesetz) see Vogel 2012a.
[6] Because of inconsistent formats we could only analyse corpus B and C.

determinative element completes and clears that frame with an attribution. The ten most frequent compounds with *arbeitnehm* as determinandum are the following:

Table 3: Compounds with *arbeitnehm* as determinandum

Nr.	f	Kompositum (Type)	[English translation]
1	851	*Leiharbeitnehmer*	agency staff
2	371	*Wanderarbeitnehmer*	migrant labourer
3	125	*Teilzeitarbeitnehmer*	part-time employee
4	99	*Leiharbeitnehmerin*	female agency workers
5	66	*Vollzeitarbeitnehmer*	full-time employee
6	51	*Stammarbeitnehmer*	permanent employee
7	18	*Altersteilzeitarbeitnehmer*	partial retirement employee
8	16	*Fremdarbeitnehmer*	interim / subcontracted employees
9	14	*Saisonarbeitnehmer*	seasonal labourer
10	14	*Gesamtarbeitnehmer*[*-vertretung*]	General works council

Compounds like these ones are very concentrated marks of legal case groups and recurrent groups of persons. If we explore particular compounds in a diachronic way we can see that, for example, legal proceedings about temporary agency workers (*Leiharbeitnehmer*) increase.

Fig. 4: *Leiharbeitnehmer* at the corpus (screenshot of the concordance plot of AntConc)

Compounds with *arbeitnehm* as a determinative element display how judges try to organize different groups of cases as prototypes. They write about

(1) *Arbeitnehmerähnlichkeit* (‚similarity of employee'), *Arbeitnehmereigenschaft* (attributes of employee), *Arbeitnehmerbegriff* (meaning of employee), *Arbeitnehmerstatus* (status of emlployee), [*fiktiver*] *Vergleichsarbeitnehmer* (‚comparitive employee'), *Arbeitnehmerkategorien* (categories of employee), *arbeitnehmertypisch* (employee-typical)

In this perspective of prototyping a new category called *arbeitnehmerähnliche Person* (referring to § 5 Abs. 1 S. 2 ArbGG) was born. *Arbeitnehmerähnliche Personen* are ›persons, selling their services to someone paying for those services, but these persons have far lesser degree of temporarily or locally self-determination as prototypical employees‹.

(b) If we search for predication phrases like *arbeitnehm is* Y, we get legal efforts to find and to reproduce definitions, created by judges.

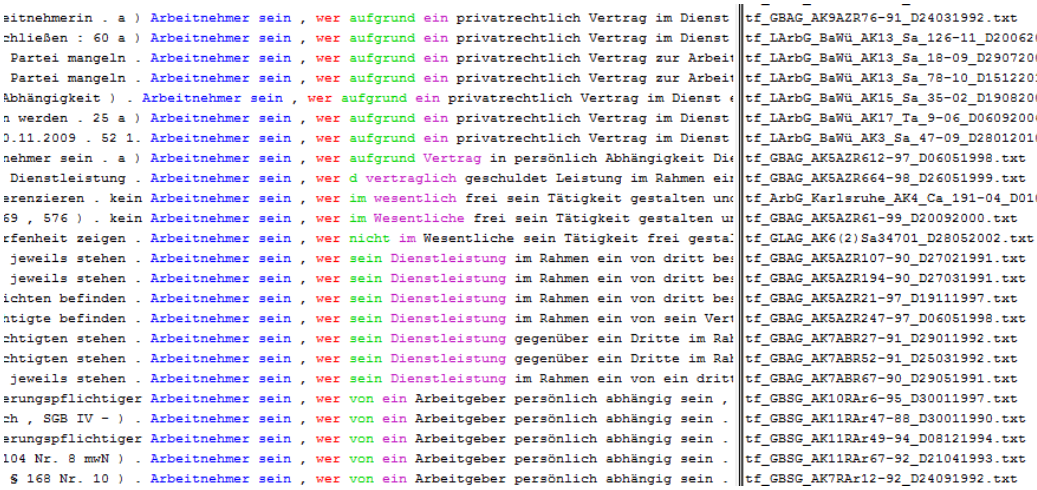


Fig. 5: Definitions for *arbeitnehm*, created by judges (Screenshot of AntConc)

We cannot only see one definition, but competing definitions. One of the most frequent patterns is the following:

(2) ***Arbeitnehmer ist, wer*** *seine Dienstleistung im Rahmen einer von Dritten bestimmten Arbeitsorganisation erbringen muss. Die Eingliederung in die fremde Arbeitsorganisation wird dadurch besonders deutlich, dass ein Arbeitnehmer hinsichtlich Zeit, Dauer und Ort der Ausführung des übernommenen Dienstes einem umfassenden Weisungsrecht des Arbeitgebers unterliegt. Häufig tritt auch eine fachliche Weisungsgebundenheit hinzu.* [Gefolgt von zahlreichen Quellenangaben mit Verweis auf das BAG]

Translation: ›An employee is someone, who has to provide his services in an organisational structure determined by others regarding time duration and place of services. Often an employee is also bound by professional instructions.‹ [followed by numerous references to decisions of the same federal labour court.]

(c) Co-occurrences of the expression *arbeitnehm* refer to broader context relationships. To analyse these relationships we collected all expressions given 15 words left to 15 words right of our start expression and filtered out these expressions, which are significant in a statistical sense[7]. To cluster the resulting words into different semantic fields we explored the particular (con-)texts in the concordance view and with co-occurrence analysis of secondary or tertiary rank, that is systematic context exploration through co-occurrences of co-occurrences and so on. To get an overview of the data structure we also used visualization methods like network analysis (here: by the help of Gephi[8]):

[7] That means that we do not only used a frequency sorted word list but statistical tests (Chi Square) that give information about the specific of the collected co-occurrences depending on the corpus population and levels of contingency or statistical expectations.

[8] http://gephi.org/, The Open Graph Viz Platform (12.07.2013).

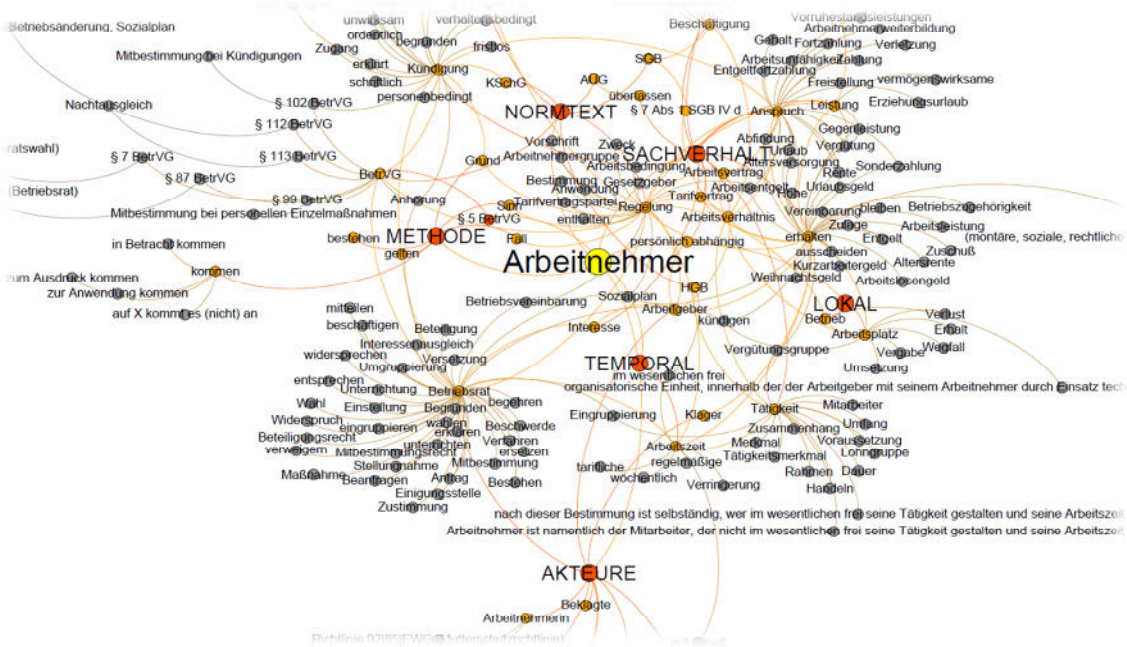


Fig. 6: Network analysis: Attribution network of *Arbeitnehmer* (Screenshot of Gephi)

The evaluation of the co-occurrence clusters focuses on important parts of the *Arbeitnehmer*-associated frame, for example:

- Relevant norm texts (laws) and legal domains attributing the *Arbeitnehmer*-Frame: paragraphs of *Betriebsverfassungsgesetz* (BetrVerfG) regarding to questions of employee participation and collective protection; German Civil Code (*Bürgerliches Gesetzbuch*, BGB) regarding the relationship between 'service seller' and 'service consumer' and others. Moreover, you can see the methodology judges' use, for example a high degree of auto-referentiality (self-citation): BAG ⇨ BAG; BSG ⇨ BSG and so on. Of course, it would be interesting analysing the full network of citations and the different types of authorities, but that would lead too far.
- Recurrent actors or stakeholders struggling for the *Arbeitnehmer*-Frame: complainer (*Kläger*) and defendant (*Beklagter*), worker's council (*Betriebsrat*) and worker's union (*Gewerkschaft*), employer (*Arbeitgeber*), employee (*Arbeitnehmer*) and the role of female employee (*Arbeitnehmerin*) especially in the last 20 years and framed by European law[9]. We have to point out, that the meaning of *employer* (*Arbeitgeber*) only is made clear by the negation of employee (*Arbeitnehmer*). A recurrent pattern is: 'employer is who employs an employer' (*Arbeitgeber ist, wer Arbeitnehmer beschäftigt*)[10]. Furthermore, many co-occurrences refer to different case groups, for example distinguishing gender (*männlich/weiblich* [male/female]), age (*jung/alt* [young/old])[11], duration and period of work (*Vollzeit* [fulltime], *Teilzeit* [parttime], *vorübergehend* [temporary], *wöchentlich* [weekly] and so on), healthy (*arbeitsunfähig* [disabled], *krank* [ill], *Prognose* [health prediction]) and others.
- Circumstances and correlating actions framing the *Arbeitnehmer*: The most important point is the expression *Tätigkeit* ('action / employment') because of its role in distinguishing employee (*Arbeitnehmer*) and freelancer (*Selbstständiger*). Regarding to § 84 Abs. 1 Satz 2, Abs. 2 HGB there are two patterns, constituting the employee and freelancer as antonyms:

[9] Art. 8, 11 of 92/85/EWG, *Mutterschutzrichtlinie* and Art. 2 of 76/207/EWG, *Gleichbehandlungsgrundsatz*.
[10] The German legal expression *Beschäftigter* (*employer*) (social law) is quasi synonymic with the expression *Arbeitnehmer* (labour law).
[11] Cases with young employees increased in the last years because of financial crises and youth unemployment.

(3) *Arbeitnehmer ist namentlich der Mitarbeiter, der nicht im Wesentlichen frei seine **Tätigkeit** gestalten und seine Arbeitszeit bestimmen kann.* (Employee is a person, who cannot arrange its employment and its working time essentially independent.)

(4) [Eine Person] *ist selbständig, wer im Wesentlichen frei seine **Tätigkeit** gestalten und seine Arbeitszeit bestimmen kann.* (Freelancer is a person, who can arrange its employment and its working time essentially independent.)

– Another symbol for the antonymic relationship is the significant co-occurrence *oder* (or) with its disjunctive meaning: *eine abhängige Beschäftigung oder eine selbständige Tätigkeit* (dependent employment or freelanced engagement).

(d) These examples should be enough to demonstrate the method, using corpus linguistic methods to structure big data of legal texts. Now I will try to summarize the results of the particular pattern analysis to describe our hypotheses bottom up about the legal *Arbeitnehmer* concept.

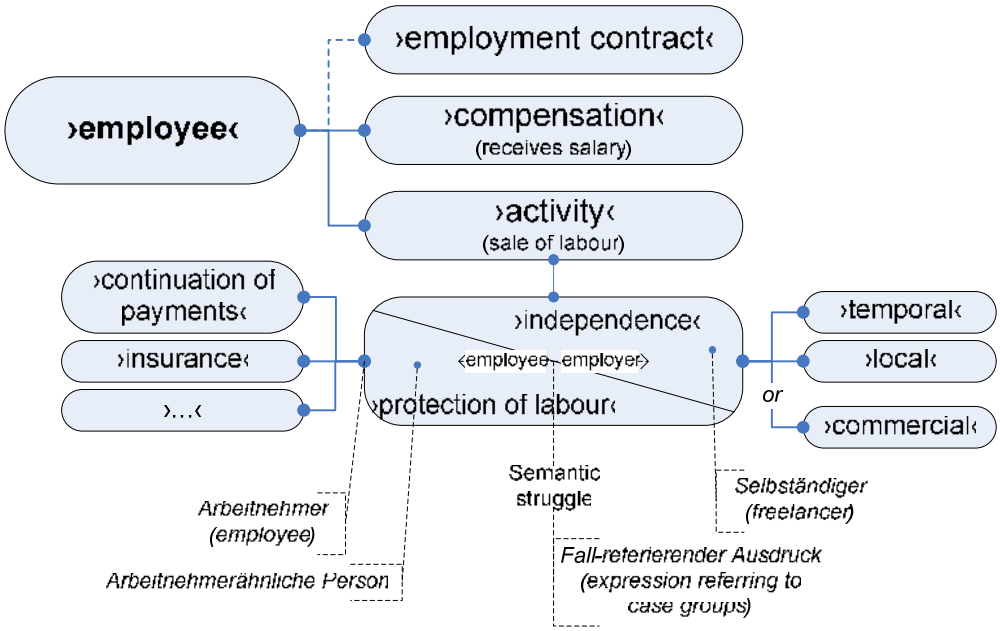


Fig. 7: Semantic structure of *employee* in German labour law dogmatic

The semantic structure of ›employee‹ in German labour law dogmatic contains four essential frame slots. The different concretions or fillers (as recurrent expression patterns) of these slots constitute particular groups of persons with different rights and duties. Generally it's a question of ›persons‹ selling their services (first slot) and being paid for it (second slot). This relationship can be, but doesn't have to be formalized as an employment contract (third slot). The fourth slot classifies the circumstances of the offered services within two dimensions, which are the ›degree of independence‹ (regarding to temporal, local or commercial aspects) and the ›degree of protection of labour‹ (for example continuation of payments, insurance, dismissals protection and so on). The different expressions and speech patterns negotiated in legal discourses refer to different prototypical combinations of slot fillers. A person called *Arbeitnehmer* (employee) has prototypically no self-determination and therefore full protection of labour. By contrast persons called *selbständig*

(freelancers) are totally independent, but have no protection of labour. If we observe the negotiation strategies of the struggling stakeholders, we can see that employees usually illustrate their work as bounded by instructions whereas employers emphasize the independence. We can describe labour law as a discourse (in the sense of Michel Foucault 1974a, 1974b: 133), where employers always constitute new types of as-it-were-freelancers ('Als-ob-Selbstständige') and German labour legislation often is late to react (cf. so called *Werkverträge* or *Scheinselbstständigkeit* / ostensible self-employment).

5. Conclusion

This contribution tried to introduce to the methodological approach of legal corpus pragmatics, using the example of the concept *employee* in German labour law. Finally I want to discuss possible drawbacks and opportunities of computer assisted studies of legal language and discourse:

1. On the one hand, corpus linguistics, especially computer based algorithms cannot calculate legal meanings. Interpretation, that is contextualization of expressions using other sensory input and background knowledge, is and will always be a challenge to human beings. Automates or machines can only connect predefined information, but they cannot evaluate this information in a context of struggling interests and the general relativity of ways to describe the world. Given this reason, attempts to substitute lawyers with computers as „subsumption automats"[12] will always fail. The computer cannot replace the judge; it cannot decide (Kotsoglou 2014).
2. On the other hand, that doesn't mean that computer assisted analysis methods are useless. The computer can help us to structure big text data – large amounts of texts – to contrast or control our introspection and – in the words of the lawyer Ralph Christensen (Mannheim) – "lawyer's impressionism". The meaning of a word is its use in the language, but we cannot remember all or at least all so called 'most important' examples of word use. Even dictionaries often fail (cf. Mouritsen 2010)[13]. Algorithms like co-occurrence analysis are able to support our memory, especially in a digitalized world of growing legal texts and promulgation through media of law. That is to say, legal corpus pragmatics provide a method for 'computer assisted reading' of law.
3. Our pilot studies illustrate that corpus assisted interpretation methods also work well in legal contexts. In the United States we can already find first court decisions using corpus linguistics[14] and existing text corpora. However, we don't have adequate legal text corpora yet. That's why I have started a project with the duration time from 2014 until 2017, supported by the Heidelberg Academy of Science and Humanities, to develop a legal reference text corpus (*Juristisches Referenzkorpus*, *JuReko* – www.jureko.de) together with Dr. iur. Hanjo Hamann (Bonn). This corpus shall contain thousands of court decisions, texts of legal scholars, commentaries and laws and shall be assembled in the next three years (cf. http://www.jureko.de; cf. Vogel/Hamann 2015).

[12] The judge as non-subjective ‚robot', cf. from Rave et al (Ed.) 1971 until Raabe et al 2012.

[13] Dictionaries are not always a proper source of arguments in a debate about the meaning of a legal clause. Why not? – The first problem is that dictionaries are normally not based upon legal texts, but on different, non-legal speech varieties. Second problem, "the dictionary is not a fortress": As Stephen Mouritsen (2010) pointed out, dictionaries only try to describe the usage in the past, but they do not *prescribe* it. Dictionaries are not 'laws on language use'.

[14] Cf. the decision of the Supreme Court of State of Utah, 2011 UT 38, 266 P.3d 702 (available under: http://www.utcourts.gov/opinions/supopin/InReEZ071911.pdf, 04.06.2014). See also: "U.S. Supreme Court uses corpus created by BYU professor Mark Davies" (Deseret News, 13.03.2011; available under http://www.deseretnews.com/article/700118257/US-Supreme-Court-uses-corpus-created-by-BYU-professor-Mark-Davies.html?pg=all).

4. With a legal reference corpus we will be able to describe and to explore legal language and legal discourse in more detail, especially regarding legal vocabulary, citation networks, the most common arguments, rules of establishing authority, the empirical status of standards and indefinite legal concepts and other questions. Furthermore, together with linguists and lawyers of the Federal Ministry of Justice (Germany) we also explore corpus assisted methods to optimize legislation. If you create a new text of law you have to anticipate and to connect with future legal language usage. But often it is not clear enough whether you should create a new word or you should better use an existing and established speech pattern associated with the desired concept to prevent misunderstandings.
5. Last but not least I think we need more interdisciplinary interactions between traditional legal methodology, legal linguistics and corpus linguistics, at the beginning of legal education. Bringing these approaches together would be a contribution to a theory of practice as practice (Bourdieu et al 2009) with sustainable long-term benefits.

6. References

Anthony, Laurence. 2005. AntConc: design and development of a freeware corpus analysis toolkit for the technical writing classroom. In *Professional Communication Conference*. *IPCC 2005. Proceedings*, 729–737.

Auer, Peter. 1986. Kontextualisierung. In *Studium Linguistik* (19), 22–47.

Barsalou, Lawrence W. 1992. Frames, concepts and conceptual fields. In Lehrer, Adrienne & Eva Feder Kittay (eds.), *Frames, fields and contrasts. New essays in semantic and lexical organization*., 21–74. Hillsdale: N.J-L. Erlbaum Associates

Best, Karl-Heinz. 2000. Unser Wortschatz. Sprachstatistische Untersuchungen. In Eichhoff-Cyrus, Karin M. & Rudolf Hoberg (eds.), *Die deutsche Sprache zur Jahrtausendwende. Sprachkultur oder Sprachverfall?,* 35–52. Mannheim: Dudenverl (Duden, 1).

Best, Karl-Heinz. 2006. *Quantitative Linguistik. Eine Annäherung*, 3. Göttingen: Peust & Gutschmidt.

Bourdieu, Pierre, Cordula Pialoux & Bernd Schwibs. 2009. *Entwurf einer Theorie der Praxis. Auf der ethnologischen Grundlage der kabylischen Gesellschaft*. 291. Frankfurt am Main: Suhrkamp.

Busse, Dietrich. 1992. *Textinterpretation. Sprachtheoretische Grundlagen einer explikativen Semantik*. Opladen: Westdeutscher Verl.

Felder, Ekkehard, Marcus Müller & Friedemann Vogel (eds.). 2012. *Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analysen*. Berlin: De Gruyter.

Foucault, Michel. 1974a. *Die Ordnung des Diskurses*. Frankfurt am Main: Suhrkamp.

Foucault, Michel. 1974b. *Die Ordnung der Dinge. Eine Archäologie der Humanwissenschaften*. Frankfurt am Main: Suhrkamp.

Goffman, Erving. 1974. *Frame analysis. An essay on the organization of experience*. Illinois: Harper & Row.

Goffman, Erving. 2002 [1967]. *Interaktionsrituale: über Verhalten in direkter Kommunikation*. Frankfurt am Main: Suhrkamp.

Gumperz, John Joseph. 1982. *Discourse strategies*. Cambridge: University Press.

Hörmann, Hans. 1980. Der Vorgang des Verstehens. In Kühlwein, Wolfgang (ed.): *Sprache und Verstehen*, 17–29. Tübingen: Narr.

Kotsoglou, Kyriakos N. 2014. Subsumtionsautomat 2.0. Über die (Un-)Möglichkeit einer Algorithmisierung der Rechtserzeugung. *Juristenzeitung* 69 (9), 451–457.

Kudlich, Hans & Ralph Christensen. 2004. Die Kanones der Auslegung als Hilfsmittel für die Entscheidung von Bedeutungskonflikten. *Juristische Arbeitsblätter*, 74–83.

Minsky, Marvin. 1975. A framework for representing knowledge. In Winston, Patrick Henry & Berthold Horn (eds.), *The psychology of computer vision,* 211–277. New York: McGraw-Hill.

Morlok, Martin. 2004. Der Text hinter dem Text. Intertextualität im Recht. In Blankenagel, Alexander, Ingolf Pernice & Markus Kotzur (eds.), *Verfassung im Diskurs der Welt. Liber Amicorum für Peter Häberle zum siebzigsten Geburtstag. Unter Mitarbeit von Peter Häberle,* 93–136. Tübingen: Mohr Siebeck.

Morlok, Martin . In press. Intertextualität und Hypertextualität im Recht. In Vogel, Friedemann (ed.): *Zugänge zur Rechtssemantik. Interdisziplinäre Ansätze im Zeitalter der Mediatisierung zwischen Introspektion und Automaten*. Berlin & New York: Walter de Gruyter.

Mouritsen, Stephen C. 2010. The Dictionary Is Not a Fortress: Definitional Fallacies and a Corpus-Based Approach to Plain Meaning. In *Brigham Young University Law Review*, S. 1915–1980

Müller, Friedrich, Ralph Christensen & Michael Sokolowski. 1997. *Rechtstext und Textarbeit*. Berlin: Duncker & Humblot.

Nelson, Todd D. 2006. *The psychology of prejudice*. Boston, Mass: Pearson Education.

Raabe, Oliver, Richard Wacker, Daniel Oberle, Christian Baumann & Christian Funk. 2012. *Recht ex machina. Formalisierung des Rechts im Internet der Dienste*. Berlin & Heidelberg: Springer Vieweg.

Rave, Dieter, Hans Brinkmann & Klaus Grimmer (eds.). 1971. *Paraphrasen juristischer Texte*. Darmstadt: Dt. Rechenzentrum.

Schützeichel, Rainer. 2007. Soziale Kognitionen. In Schützeichel, Rainer (ed.). *Handbuch Wissenssoziologie und Wissensforschung*. 433–449. Konstanz: UVK-Verlags-Gesellschaft.

Starck, Fritz & Roland Deutsch. 2002. Urteilsheuristiken. In Frey, Dieter & Martin Irle (eds.), *Theorien der Sozialpsychologie. Motivation und Informationsverarbeitung,* 352–385. Bern: Huber.

van Dijk, Teun Adrianus. 1999. Context Models in Discourse Processing. In van Oostendorp, Herre & Susan R. Goldman (eds.), *The construction of mental representations during reading,* 124–148. Mahwah, NJ: Erlbaum.

Vogel, Friedemann. 2012a. *Linguistik rechtlicher Normgenese. Theorie der Rechtsnormdiskursivität am Beispiel der Online-Durchsuchung*. Berlin: De Gruyter.

Vogel, Friedemann. 2012b. Das Recht im Text. Rechtssprachlicher Usus in korpuslinguistischer Perspektive. In Felder, Ekkehard, Marcus Müller & Friedemann Vogel (eds.), *Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analysen*, 314–353. Berlin: De Gruyter.

Vogel, Friedemann. 2012c. Das LDA-Toolkit. Korpuslinguistisches Analyseinstrument für kontrastive Diskurs- und Imageanalysen in Forschung und Lehre. *Zeitschrift für angewandte Linguistik* 57 (1), 129–165.

Vogel, Friedemann (ed.). In press. *Zugänge zur Rechtssemantik. Interdisziplinäre Ansätze im Zeitalter der Mediatisierung zwischen Introspektion und Automaten*. Berlin & New York: De Gruyter.

Vogel, Friedemann & Hanjo Hamann. 2015. Vom corpus iuris zu den corpora iurum – Konzeption und Erschließung eines juristischen Referenzkorpus (JuReko). In Heidelberger Akademie der Wissenschaften (eds.), *Jahrbuch der Heidelberger Akademie der Wissenschaften für 2014*. Heidelberg: Winter.

Dr. Friedemann Vogel has worked as a professor of media linguistics at the Institute of Media Culture Science at University of Freiburg, Germany, since 2012. In his PhD (2008) he explored the genesis of legal norms in the context of legislation, published 2012 (Mouton de Gruyter). Vogel has developed an approach to combine methods of computer mediated linguistics and pragmatics to analyse legal norms, first time in Europe. Together with E. Felder (Heidelberg) he prepares an European Handbook of Legal Linguistics (forthcoming). Further fields of interests: corpus linguistics and programming, discourse studies, stereotype analysis, language and democracy, computer mediated communication and conversational analysis (focus on conflict settings). Different fellowships at Universities of Beijing (China), Budapest (Hungary), Zurich (Switzerland), Heidelberg, Regensburg and Freiburg (Germany) and the Academy of Science and Humanities (Heidelberg).

Prof. Dr. Friedemann Vogel
Albert-Ludwigs-Universität Freiburg
Institut für Medienkulturwissenschaft
Werthmannstraße 16
79085 Freiburg im Breisgau (Germany)
Telefon: +49-(0)-761-203 97845
Fax: +49-(0)-761-203 97846
Web: www.friedemann-vogel.de