Stefan Engelberg, Alexander Koplenig, Kristel Proost, Edeltraud Winkler

Argument structure and text genre: cross-corpus evaluation of the distributional characteristics of argument structure realizations

1. Introduction

Quantitative investigations into the distributional properties of argument structures have led to insights into the relationship between verbs, the argument structures they occur in, and the semantic verb classes they belong to. These studies have revealed how particular verbs show idiosyncratic preferences for particular argument structures and how particular argument structures attract particular verbs (Stefanowitsch/Gries 2003; Gries/Stefanowitsch 2004, 2010). From the perspective of Construction Grammar, these data not only point to the existence of relationships, including inheritance relationships, between constructions, but also provide the empirical basis for the discussion of identity conditions for the establishment of constructions.[1] They are also a prerequisite for addressing issues such as the productivity of constructions (e.g., Barðdal 2008), the mechanisms underlying the extension of a construction to new lexical items (cf. e.g., Boas 2011), and the attraction between lexical elements filling different slots in argument structure constructions (e.g. Engelberg et al. 2011).

From a perspective more closely connected to valency theory, frequency data on verbs and valency patterns have revealed that verbs show certain preferences with respect to their valency specifications and valency alternations (Köhler 2005, Engelberg 2009, Cosma/Engelberg 2013). This also means that valency lexicography is confronted with a growing wealth of new quantitative data that might influence the content and structure of valency

[1] Cf. the debate in Welke (2009), Engelberg (2009), and Goschler (2011) about whether German verbs of sound emission used as movement verbs establish a particular construction and how this construction might relate to a general movement verb construction.

DOI 10.1515/lexi.2012–0004

dictionaries in the future. Frequency data have also been used to automatically induce verb classes from the distribution of valency frames (cf. Schulte im Walde 2003, 2009).With respect to the ongoing debate about the relationship between the concepts of valency and construction (cf. Jacobs 2009, Welke 2009, Boas 2010, Stefanowitsch 2011, Engelberg et al. 2011), quantitative data on argument structures might prove particularly interesting.

From a general cognitive viewpoint, quantitative data on argument structures and related phenomena have also been used in the debate about the role of the frequency of a linguistic structure which has been argued to be crucial to its entrenchment in the language system (e.g., Bybee/Beckner 2010, Schmid 2010, Gries 2010). Quantitative data of this type have also shed light on the principles of the diachronic change of constructions (e.g., Bybee 2010) and on the acquisition of verbs and syntactic structures (e.g., Tomasello 2003, 2006; Behrens 2009, 2011).

Since the investigations mentioned cover a wide range of different issues, it is surprising that only a few studies have been concerned with the extent to which corpus choice influences the results of quantitative investigations. Gries (2006, 113) claims that, generally, "there is not much work that systematically explores the issues of variability within corpora (i.e. corpus homogeneity) and between corpora." This holds in particular for investigations into argument structure and might be due to the fact that many corpora focus on newspaper texts, are of insufficient size, or do not have the annotations that would have facilitated the search for particular argument structures. Even if corpora are richly annotated, the subtle semantic indicators of a particular argument type often require time-consuming manual processing of the data. While manual processing of the data is problematic with respect to a single corpus, it is even more so with respect to a study involving different corpora. For these reasons, the question of how a particular corpus and the kinds of texts it contains affects the quantitative properties of argument structures is not an easy one to answer.

German valency theory has occasionally noticed the influence of text genre on the choice of valency patterns. However, systematic empirical studies are rare. Helbig (1985) observed that different text genres make different use of possible verb valency patterns, and Sommerfeld (1999) studied the connection between valency and the development of text genres. A small number of papers investigated particular text genres with respect to valency patterns. Sommerfeldt (1993) looked at sports reports, public notices, and instruction manuals; Schwittalla (1985) – in a small quantitative study – investigated obituaries, and Schatte (2002) worked on valency and special language from a lexicographical point of view.

In the present paper, we will present a sample-based comparison of six corpora with respect to the distribution of the argument structures of verbs. In particular, we will show (i) how the type of corpus (newspaper, fiction, spoken language, scientific texts, etc.) systematically influences the distributional patterns of argument structures, (ii) whether the degree of influence on these distributional patterns depends on the verb class under investigation, and (iii) to what extent the cross-corpus distributional differences observed can be accounted for by genuine properties of particular text genres.

The paper proceeds as follows. In section 2, we will present a brief overview of related investigations. Section 3 describes the design of our study, and section 4 its results. In section 5, we will discuss the results and interpret them with respect to the properties of text genres, verb classes, and argument structure realization patterns.

2. Some studies on variation across corpora

Some previous quantitative investigations have concentrated either on verb subcategorization frequencies in different corpora or on the variability between and within corpora. These studies make use of different methods, analysis tools, and experiments. In this section, we will look more closely at some of these.

Roland/Jurafsky (1998) analyze how verb subcategorization frequencies are affected by corpus choice. They discovered that verb sense on the one hand and discourse type on the other are responsible for the different subcategorization frequencies of verbs in different corpora. Discourse influences proved to result from the way in which verb use is affected by the different discourse types. These differences are to some extent predictable from the discourse types that occur in a particular corpus. The semantic influences appeared to be based to a large extent on the fact that different corpora often use different verb senses, which show different subcategorization frequencies. That is, the semantic context of the discourse is associated with certain verb senses. These differences are hardly predictable from the relative frequencies of each of the possible senses of the verb in a corpus. Furthermore, there appeared to be significant differences between the verb subcategorization frequencies generated through experimental methods (in psycholinguistic studies) and those observed from the use of corpus methods.

The role of verb sense and the verb subcategorization frequency differences between corpora are the central themes dealt with by Roland et al. (2000). The aim of this study was to obtain stable cross-corpus subcategorization probabilities to provide norming data for psychological experiments. Most of the verbs analyzed show remarkably stable subcategorization preferences between British and American corpora as well as between balanced corpora and financial news corpora. Where the verbs show differences, these shifts in subcategorization are largely the result of subtle verb sense differences between the genres present in each corpus. This observation suggests that stable cross-corpus subcategorization frequencies may be found when verb sense is adequately controlled. In the authors' opinion, it might be possible to use verb frequencies and subcategorization probabilities of multi-sense verbs for measuring the degree of difference between corpora. Like Roland et al. (2000), Gahl/Jurafsky/Roland (2004) set out to collect useful data for norming behavioral experiments. Their aim was to gather the subcategorization frequencies for a larger number of verbs than had been considered before, and they based their study on a corpus larger than those used before. They used British English as well as American English corpus data and compared methods and corpora. Their norms were accompanied by an explicit coding manual.

Roland's dissertation (2001) also deals with verb senses and verb subcategorization probabilities. He presents a model that could make verb subcategorization predictions based on the semantic context that precedes the verb in corpus data. The results of this procedure are the same as the predictions that are made by human subjects given the same contexts. Evidence presented in this work could show that important causes of the subcategorization frequency differences between corpora are different senses of verbs and their corresponding differences in subcategorization as well as inherent differences between the production of sentences in psychological experiments and language use in context. For this reason, verb subcategorization probabilities should be based on individual senses of verbs rather than the whole verb lexeme.

Schlüter (2006) examines the reliability of the results of individual corpus linguistic analyses by comparing them to studies covering similar or identical ground. As the subject of investigation, he chooses the present perfect, and he compares ten studies dealing with the main aspects of the present perfect. These studies were based on different language corpora and applied different software tools to examine the data. The comparison shows that the results of the ten studies exhibit the same general tendencies. Some inconsistencies are inevitable, but they have a stronger effect on studies based on smaller corpora. Since larger corpora contain a greater variety of text types, the results of the studies based on large corpora can be considered to be more reliable.

Some more recent work (e.g., Gries 2006 and 2011) makes use of the methods of collostructional analysis. Gries (2006) states that the variability of the data used is a key issue when interpreting and comparing the results between different studies. Corpora are variable internally as well as externally. Gries is mainly concerned with how to identify and quantify the degree of variation in the results, how to investigate the source of the observed variation, and how to find out the degree of homogeneity of a corpus with respect to a particular phenomenon. He shows that corpus variability involves making decisions concerning the parameter of interest and the desired level of granularity. It is also worth noting that quantitatively different results do not necessarily yield qualitatively different theoretical conclusions. Gries (2011) investigates methodological issues associated with the use of corpus data, especially the degree of granularity providing the most insightful results. He investigates two granularity parameters: (i) inflectional-form-based vs. lemma-based corpus analyses and (ii) register variation. One of the results is again that not all quantitative distinctions are correlated with meaningful differences from a linguistic point of view. However, differences in register produced larger quantitative differences than the distinction between inflectional-form-based and lemma-based assessment of argument structure data.

In general, earlier studies have shown that different corpora, in particular those reflecting different text genres, yield different frequencies for argument structure phenomena. Gries claims that his method of collexeme analysis is not affected much by the cross-genre differences observed. Since collostructional analyses do not form part of our study, we will not be dealing with this issue. Roland, Jurafsky, and colleagues put a lot of emphasis on verb senses; particular text genres or discourse types are associated with particular verb senses which in turn pattern with particular subcategorization frames. Since in our own study we consider all verbs under investigation to be monosemous, the frequency differences found in the investigation are related to corpus type/text genre only.[2]

---

[2] As in our study, Roland et al. (2000, 30) "used a broadly defined notion of sense rather than the more narrowly defined word senses used in some on-line word sense resources such as WordNet."

3. Methodology

3.1 Starting point

The starting point for the present study is the research project *Verben und Argumentstrukturen* carried out at the Institut für Deutsche Sprache, which deals with the argument structure patterns of German verbs (cf. the contributions in Winkler 2009). On a descriptive level, the project sets out argument structure patterns and their semantic and syntactic properties and captures the predictable or idiosyncratic behavior of verbs with respect to their occurrence in these argument structure patterns. On a theoretical level, the project aims at a critical evaluation of (valency-based) projectionist language theories versus construction-based theories. The quantitative and qualitative corpus-based investigations currently being carried out point to a net of fine-grained argument structure patterns connected by Wittgenstein-type family relationships that interacts with a high number of idiosyncratic lexical specifications (Engelberg et al. 2011).

While gathering quantitative data for this investigation, we noticed that some of the quantitative results obtained for some verbs depended on the textual nature of the main corpus which consists mainly of newspaper texts. We, therefore, decided to validate our results by repeating the analysis with samples from other corpora representing different text genres. We expected the results of this investigation not only to reveal interesting differences in the cross-corpora behavior of the verbs investigated but also to provide general indicators (e.g., verb class) for investigations into the use of valency patterns and argument structure that would allow us to predict whether or not a verb is likely to show genre-specific behavior.

3.2 Data Collection

For this cross-corpus investigation, we chose six corpora.[3] Even though we are not claiming that each of these corpora represents one particular genre, we assume that different genres are associated with each of five of the six corpora. Two corpora (FICTION1, FICTION2) are assumed to represent more or less the same genres:[4]

- ***NEWSPAPER (IDS)***: virtual corpus from DeReKo containing mainly newspaper texts (90%), some other non-fiction texts, and some fiction.
- ***FICTION1 (IDS)***: virtual corpus from DeReKo containing general fiction (mainly novels, e.g., Bichsel, Grass, Lenz, Thomas Mann, Walser; light fiction and autobiographical texts, e.g., Klemperer).

---

3 We are grateful to Agata Sokolowski for collecting and analyzing much of the data used in this study, and we thank Peter Meyer for valuable comments regarding an earlier version of this paper.

4 The "DWDS-Kernkorpus", located at the "Berlin-Brandenburgische Akademie der Wissenschaften," is a corpus of 100 million running words equally distributed over the ten decades of the 20th century and four sets of text genres (fiction, general non-fiction, science, newspapers); cf. Geyken (2007) for more on the content of the corpus; DeReKo, the German Reference Corpus, located at the "Institut für Deutsche Sprache" (IDS) contains more than 4,000 million running words with a strong focus on texts from newspapers and journals.

- ***Spoken (IDS)***: corpus of spoken language ("Deutsch heute") located at the IDS containing recordings/transcripts of people giving directions as well as interviews with pupils and students who were questioned about their linguistic biographies; media corpus ("Medien") at the IDS comprising different kinds of texts (news broadcasts, sports reports, talk shows, etc.) broadcast on TV.
- ***Fiction2 (DWDS)***: part of the "DWDS-Kernkorpus" that contains general fiction, mainly novels (e.g., fiction by Anders, Böll, Degenhardt, Dürrenmatt, Enzensberger, Johnson, Koeppen, Walser, Wellershoff).
- ***Non-Fiction (DWDS)***: part of the "DWDS-Kernkorpus" that contains non-fiction texts (e.g., biographies, guides to etiquette, letters, theater programs, self-help books, prescription drug information, cookbooks).
- ***Science (DWDS)***: part of the "DWDS-Kernkorpus" that contains scientific texts (e.g. on biology, economics, linguistics, musicology, pedagogy, philosophy, political science, science of art, sociology, theology).

We investigated the distribution of argument realization patterns for 16 verbs from five semantic verb classes:

A) ***Psych-verbs*** (*freuen* 'become/make happy', *wundern* 'be astonished/astonish', *ärgern* 'get/make angry') denote a relation between an experiencer *x* and a stimulus *p* (essentially a proposition-like entity).
B) ***Connective verbs*** (*widersprechen* 'contradict', *erklären* 'explain', *verursachen* 'cause') denote relations between two proposition-like entities; they also allow the realization of NPs expressing human participants; for example, *widersprechen* 'contradict' expresses a relation between a proposition *p*, uttered by a participant *x*, and a proposition *q*, held by a participant *y*.
C) ***Directed emotion verbs*** (*lieben* 'love', *hassen* 'hate', *bewundern* 'admire') denote emotions between an animate experiencer *x* and a target of emotion *y* (an animate being, an object, or a proposition).
D) ***Perception verbs*** (*empfinden* 'feel/sense', *fühlen* 'feel', *hören* 'hear') describe a relation between an animate participant *x* and the participant (or event/proposition-like entity) *y* that *x* experiences or becomes cognitively aware of.
E) ***Action verbs*** (*arbeiten* 'work', *bauen* 'build', *kochen* 'cook', *malen* 'paint') denote a (mostly) physical action of medium complexity performed by an agent *x* with respect to an object *y*.

As mentioned above, most of the sentences sampled from the Newspaper (IDS) corpus were already annotated in our research project. For the rest of the corpora, random samples (N = 100) were initially drawn for each verb, and these were then manually cleared from all irrelevant sentences.[5] In order to arrive at a data set of at least 100 sentences from each

---

[5] This step eliminated all sentences that were either inappropriate or incomplete. Therefore, examples (i) for the verb *malen* 'to paint'and (ii) for the verb *veursachen* 'to cause' were excluded from the annotation.
(i) Sie ist **mal** wieder viel zu dünn angezogen für diese Jahreszeit und friert erbärmlich.
'**Again**, she is dressed far too thinly for this time of the year and is absolutely freezing.'

corpus and each verb, we extracted another 50 sentences and repeated the manual selection of the sentences in each case. It then became clear that for some of the verbs it was not possible to arrive at a data set of 100 sentences from each corpus that could be used for the data analysis. We, therefore, decided to only include verbs in our investigation for which we were able to collect at least 20 suitable sentences from each of the six corpora and to account for the different sample sizes in the data analysis. Table 1 summarizes the resulting sample sizes.

Verb class	Verb	C1	C2	C3	C4	C5	C6	Total
A) Psych-verbs	1. *freuen* 'become/make happy'	200	100	84	94	114	54	**646**
	2. *wundern* 'be astonished/astonish'	98	100	65	98	101	49	**511**
	3. *ärgern* 'get/make angry'	200	100	68	57	76	37	**538**
	Total	**498**	**300**	**217**	**249**	**291**	**140**	**1695**
B) Connective verbs	4. *widersprechen* 'contradict'	200	200	61	91	91	96	**739**
	5. *erklären* 'explain'	94	99	77	99	100	98	**567**
	6. *verursachen* 'cause'	234	93	26	53	72	64	**542**
	Total	**528**	**392**	**164**	**243**	**263**	**258**	**1848**
C) Directed emotion verbs	7. *lieben* 'love'	100	100	99	100	100	100	**599**
	8. *hassen* 'hate'	99	100	51	98	80	57	**485**
	9. *bewundern* 'admire'	100	100	29	100	100	100	**529**
	Total	**299**	**300**	**150**	**298**	**280**	**257**	**1584**
D) Perception verbs	10. *empfinden* 'feel/sense'	150	100	94	97	100	99	**640**
	11. *fühlen* 'feel'	100	100	100	100	100	100	**600**
	12. *hören* 'hear'	100	100	100	100	100	100	**600**
	Total	**250**	**300**	**294**	**297**	**300**	**299**	**1740**
E) Action verbs	13. *arbeiten* 'work'	97	100	100	53	80	67	**497**
	14. *bauen* 'build'	100	101	96	100	100	101	**598**
	15. *kochen* 'cook'	100	100	93	100	103	98	**594**
	16. *malen* 'paint'	100	100	107	102	100	100	**609**
	Total	**397**	**401**	**396**	**355**	**383**	**366**	**2298**
Total		**1972**	**1693**	**1221**	**1442**	**1517**	**1320**	**9165**

Table 1: Sample sizes for each investigated verb (C1 = NEWSPAPER, C2 = FICTION1, C3 = SPOKEN, C4 = FICTION2, C5 = NON-FICTION, C6 = SCIENCE).

(ii) Im vergangenen Jahr ist jeder zweite durch Kinder der Altersgruppe sechs bis zehn Jahre **verursachte** Verkehrsunfall auf vorschriftswidriges Überqueren der Fahrbahn zurückzuführen gewesen.
'In the past year, half of all traffic accidents **caused** by children in the six-to-ten age group were the result of crossing the road contrary to the rules.'

The samples of corpus sentences were analyzed according to the syntactic realization of the verb's arguments. We opted for a broad, rather unspecific concept of arguments that was guided by the aim to capture all semantic roles that might show idiosyncrasies in formal realization and occurrence with respect to the particular verb or the semantic verb class they belong to. Since the number and kinds of semantic roles of this sort were not known prior to the analysis, the "argument" list for each verb was constructed as we went along starting from a broad conceptual analysis of the verb with respect to those semantic argument roles that were necessary for a lexical meaning description. The conceptual analysis for *widersprechen* 'contradict' yielded four arguments: the two propositions, Arg1 and Arg3 standing in contradiction, and the two human participants, Arg2 and Arg4 holding the opinions represented by Arg1 and Arg3, respectively.

The arguments taken into consideration define the columns of our annotation table. In turn, each occurring formal realization pattern for an argument configuration defines a row in our tables, as shown in Table 2. The occurrence of each argument realization pattern was counted while the sample sentences were analyzed.

Pattern	Arg 1	Arg 2	Arg 3	Arg 4	C1	C2	C3	C4	C5	C6
V-01		NP-nom	NP-dat		62	11	7	2	27	12
V-02		NP-nom		NP-dat	18	64	14	13	8	5
V-03		NP-nom	S-*dass*	NP-dat	1	0	2	0	0	0
V-04		NP-nom			27	42	10	29	13	1
V-05	S-dirSpeech	NP-nom			8	18	1	26	0	0
V-06	S-dirSpeech	NP-nom	NP-dat		1	0	0	0	0	0
V-07	S-dirSpeech	NP-nom		NP-dat	1	3	0	0	0	0
V-08	S-V2-subj	NP-nom			4	3	0	2	0	0
V-09	S-V2-subj	NP-nom	S-V2-subj		1	0	0	0	0	0
V-10	NP-nom		NP-dat		43	30	12	11	22	58
V-11	NP-nom-pl		Pro-dat-rec		5	3	5	1	3	3
V-12	S-inf		NP-dat		4	2	0	0	3	1
V-13	…	…	…	…	…	…	…	…	…	…

Table 2: Section of the frequency table for *widersprechen* 'contradict' (C1 = NEWSPAPER, C2 = FICTION1, C3 = SPOKEN, C4 = FICTION2, C5 = NON-FICTION, C6 = SCIENCE; S-V2 -subj: Sentence-Verb second-subjunctive, Pro-dat-rec: Pronoun-dative-reciprocal).

Examples for patterns V-02, V-07, and V-11 illustrate the distribution of argument roles in different argument realization patterns:

(1) a. [Er]$_{Arg2}$ widersprach [dem Bericht]$_{Arg3}$·
[he.NOM]$_{Arg2}$ contradicted [the report.DAT]$_{Arg3}$

b. ["Das ist unmöglich,"]$_{Arg1}$ widersprach [sie]$_{Arg2}$ [ihm]$_{Arg4}$·
["That is impossible",]$_{Arg1}$ contradicted [she.NOM]$_{Arg2}$ [him.DAT]$_{Arg4}$

c. [Die Berichte]$_{Arg1}$ widersprachen [sich]$_{Arg3}$·
[the reports.NOM]$_{Arg1}$ contradicted [themselves.DAT]$_{Arg3}$

Other verbs proved to be associated with a more complex array of roles. This is true, for example, of *malen* ('paint') whose lexical meaning was analyzed as involving the following eight roles: Arg1: the entity doing the painting, as in [*Er*] *malt gerne* ('[He] likes to paint'), Arg2: the entity painted, as in [*Die Diagonale*] *male ich aber nicht* ('But I do not paint [the diagonal]'), Arg3: the product of the act of painting, as in [*Naturalistische Bilder*] *malt sie heute keine mehr* ('Today she no longer paints [naturalistic pictures]'), Arg4: the material used for painting as in *Sie malt* [*in Öl*] ('She paints [in oils]'), Arg5: the background painted on as in [*An die Wand*] *hat jemand Sprüche gemalt* ('Someone has painted slogans [onto the wall]'), Arg6: the instrument used for painting as in *Ich male* [*mit dem Stift*] ('I draw [with a pencil]'), Arg7: the person joining the painter in painting as in *Ich male hin und wieder Bilder* [*mit meiner Enkeltochter*] ('I paint pictures now and then [with my granddaughter]'), and Arg8: the person benefitting from someone's painting something as in *Ich habe* [*ihr*] *ein paar Bilder gemalt* ('I painted [her] a couple of pictures'). We also postulated a relatively large number of semantic roles to describe the lexical meaning of *empfinden* ('feel/sense'), which may be used either as a perception verb, as in *Er empfand einen stechenden Schmerz* ('He felt a sharp pain'), an emotion verb, as in *Er empfand Mitleid* ('He felt compassion'), or as a cognition verb, as in *Er emfpand die Maßnahme als einen wichtigen Fortschritt* ('He felt the measure to be an important improvement'). To account for the different uses of *empfinden*, the following roles were taken to be associated with its lexical meaning: Arg1: the person perceiving/feeling something and Arg2: the specification of the emotion felt, as in $[Er]_{\text{Arg1}}$ *empfand* $[tiefes\ Mitleid\ mit\ den\ Opfern]_{\text{Arg2}}$ ('[He] felt [much sympathy for the victims]'), Arg3: topic and Arg4: comment, as in *Ich empfinde* $[das]_{\text{Arg3}}$ $[als\ schreckliches\ Unglück\ für\ diesen\ Platz]_{\text{Arg4}}$ ('I feel [that] to be [a terrible accident for this place]'), Arg5: the person towards whom the emotion is directed, as in *Ebenso viele Israelis empfinden ähnlich* [*für Araber*] ('Just as many Israelis feel the same [for Arabs]'), Arg6: manner, as in *Das empfinden die Politiker* [*sehr deutlich*] ('The Politicians feel that [very clearly]'), Arg7: circumstance, as in *Er empfindet seine Verwandten* [*beim Reden*] *als unwürdige Gegner* ('He feels his relatives to be unworthy opponents [while they are talking]'), and Arg8: location of perception, as in *Wir empfinden den Alkohol* [*am Zungenrand*] ('We feel the alcohol [at the edge of the tongue]'). The semantic roles postulated for *malen* and *empfinden* show that (i) apart from the well-known central roles such as agent, theme, experiencer, and stimulus, we also included more peripheral roles such as manner and circumstance for *empfinden* and co-agent for *malen* as well as roles that are not part of the verb's argument structure such as beneficiary for *malen*[6], and (ii)

---

[6] We assumed these more peripheral roles to be relevant to the meaning of some verbs but not to that of others. The role of location, for example, was taken into account only when the location in question was the place of perception (as in *Wir empfinden den Alkohol* [*am Zungenrand*]), that is, an "internal locative" in the sense of Maienborn (1996). Internal locatives appeared to be relevant to the three verbs of perception considered (*empfinden*, *fühlen,* and *hören*). However, external locatives were not taken into account for any of the verbs considered because they may occur in principle with any type of verb. Peripheral roles were also taken into account where their occurrence appeared to be related to text-genre. In the case of *empfinden*, for example, the role of manner turned out to be realized quite often in spoken language. Typical examples include *Ich empfinde das halt so* ('I happen to feel it that way'). On the whole, the distinction between central and peripheral roles assumed for the purposes of this study roughly parallels the distinction between core and non-core frame-elements in FrameNet (cf. https://framenet.icsi.berkeley.edu/fndrupal/home).

we assumed verb-specific rather than general semantic roles (e.g., 'entity doing the painting' and 'entity painted' instead of 'agent' and 'theme' for *malen* and 'person perceiving/feeling something' instead of 'experiencer' for *empfinden*).

In order to evaluate the reliability of the manual assignment, we drew a subsample of 100 sentences of our aggregated original sample. This subsample was then reannotated independently by three different coders (EW, KP, and SE) using the scheme, that is, the argument list produced by the initial annotator for each verb. For this subsample, we obtained an 83% inter-annotator agreement.[7] Only 2% of the examples were coded differently by all three annotators.

3.3 Data analysis

The complete data analysis was carried out using Stata 12 (cf. StataCorp. 2011).[8] To examine our rather exploratory research question, we first calculated bi-variate Pearson's product moment correlation coefficients between any pair of corpora for each verb to describe the similarity of the examined corpora (cf. Ludwig-Mayerhofer 2005). The bigger the value, the stronger the relationship. A high positive numerical value for two corpora x and y indicates a positive linear relationship between the frequencies of argument realization patterns for each corpus: argument realization patterns that are relatively frequent in corpus x are also relatively frequent in corpus y. It is worth pointing out that this relationship is defined as in Eq. 2:

$$r_{xy} = \frac{Cov(x,y)}{s(x)s(y)} \quad \text{(Eq. 2)}$$

where *Cov(x,y)* is the empirical covariance between the sets of observed values of x and y, while *s(x)* and *s(y)* are the standard deviations for the corresponding variables. The empirical covariance measure how two variables x and y change together: if greater values of x mainly correspond with greater values of y, it assumes positive values. It is calculated as the sum of products of the deviations of any two corresponding values x_i and y_i from their respective means $\bar{x}$ and $\bar{y}$:

$$Cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \quad \text{(Eq. 3)}$$

---

[7] Cohen's Kappa coefficient is the most widely used measure to quantify the inter-annotator agreement (cf. Carletta 1996). It is defined as

$$\varkappa = \frac{p_a - p_c}{1 - p_c} \quad \text{(Eq. 1)}$$

where p_a is the relative observed agreement between annotators, and p_c is the agreement expected by chance. However, due to the fine-grained annotation scheme constructed for each verb (cf. Table 2), the probability of chance agreement is relatively low ($p_c \approx 0$). Thus, the numerical value given in the text is roughly equivalent to κ and can be characterized as substantial.

[8] Raw data and STATA files can be obtained upon request from AK (koplenig@ids-mannheim.de).

Thus, a positive value of covariance between two corpora *x* and *y* signifies both that argument realization patterns that are above average in corpus *x* mainly correspond to argument realization patterns that are above average in corpus *y* and that argument realization patterns that are below average in corpus *x* mainly correspond to argument realization patterns that are below average in corpus *y*. However, this also implies that the covariance is scale-dependent, which in our case means that it depends on the sample sizes which are different as mentioned above (cf. Table 1). By dividing the covariance by the product of the respective standard deviations (cf. Eq. 2), we obtain a scale-independent measure ranging from -1 to 1 (cf. Ludwig-Mayerhofer 2005). This ensures that samples of different sizes can be compared in a meaningful way. All bi-variate correlation coefficients for each verb were combined in a correlation matrix (cf. Table 3).

	Newspaper	Fiction1	Spoken	Fiction2	Non-fiction	Science
Newspaper	1.00					
Fiction1	$r_{Newspaper\text{-}\ Fiction1}$	1.00				
Spoken	$r_{Newspaper\text{-}\ Spoken}$	$r_{Fiction1\text{-}Spoken}$	1.00			
Fiction2	$r_{Newspaper\text{-}\ Fiction2}$	$r_{Fiction1\text{-}Fiction2}$	$r_{Spoken\text{-}Fiction2}$	1.00		
Non-fiction	$r_{Newspaper\text{-}Non\text{-}fiction}$	$r_{Fiction1\text{-}Non\text{-}fiction}$	$r_{Spoken\text{-}Non\text{-}fiction}$	$r_{Fiction2\text{-}Non\text{-}fiction}$	1.00	
Science	$r_{Newspaper\text{-}\ Science}$	$r_{Fiction1\text{-}Science}$	$r_{Spoken\text{-}Science}$	$r_{Fiction2\text{-}Science}$	$r_{Non\text{-}fiction\text{-}Science}$	1.00

Table 3: Resulting correlation matrix for pairs of corpora.

To detect similarities between the correlation matrices, we used an agglomerative hierarchical cluster analysis (cf. Backhaus et al. 2003, 503 – 524; StataCorp 2011, 87–93). This method starts with *N* (=16) separate groups (each sized 1) for each observation (in our case, the different verbs). It then calculates a distance matrix based on the information found in the data (in our case, the respective correlation matrices). Distance is then equivalent to similarity: the closer two observations are in terms of Euclidian distance,[9] the greater the similarity between those two observations (in our case, the closer two verbs are, the more similar the distribution of argument realization patterns across corpora for those two verbs). Using the complete linkage criterion,[10] the method then proceeds by combining the closest two clusters resulting in *N-1* groups. This step is then repeated until all observations are

---

[9] The Euclidian L2-distance is calculated using the following formula:

$$L2 = \sum_{k=1}^{15} (r_{xy,i} - r_{xy,j})^2 \quad \text{(Eq. 4)}$$

Where $r_{xy,i}$ and $r_{xy,j}$ are the correlation coefficients for the k =15 corpus-pairs xy for two verbs i and j (cf. Eq. 2 and Table 3). Thus, when the distribution of argument realization patterns across corpora for two verbs is quite similar, the sum of the squared differences of the respective correlation coefficients becomes small, resulting in a small L2-distance.

[10] «Complete-linkage clustering […] uses the farthest pair of observations between the two groups to determine the similarity or dissimilarity of the two groups» (StataCorp. 2011, 88; cf. Backhaus et al. 2003, 506). Thus, at each step of the clustering, the dissimilarity between the cluster

merged into one group. This step-by-step process of clustering can be visualized in a tree diagram (a dendrogram) with the observations placed on one axis, the distance on the other axis, and U-shaped lines connecting each cluster with the height of the U representing the distance between two clusters (cf. Figure 2).

After the cluster analysis, the groups obtained (similarity clusters) were aggregated by combining the separate correlation matrices to generate one single matrix for each cluster.[11] As a next step, we conducted separate multidimensional scalings (MDS) for each of these correlation matrices. MDS is an exploratory dimension reduction technique to visualize (dis)similarities in a lower (often two-) dimensional space by preserving the higher-dimensional distances (cf. Backhaus et al. 2003, 605–672).

In other words, the MDS attempts to arrange the objects under investigation (in this case, corpora) in a two-dimensional space so that the resulting configuration plot approximates the (dis)similarities of the input matrix. It is important to point out that this also implies that the configuration is not unique.

---

formed at this step and the other clusters can be computed on the basis of the following recurrence formula:

$$d_{k(ij)} = \frac{1}{2}d_{ki} + \frac{1}{2}d_{kj} + \frac{1}{2}\left|d_{ki} - d_{kj}\right| \text{ (Eq. 5)}$$

where d_{ij} is the distance between cluster i and cluster j. $d_{k(ij)}$ is then the distance between cluster k and a newly formed cluster by combining cluster i and cluster j (Everitt et al. 2011, StataCorp. 2011, 89).

[11] It should be noted that since correlation coefficients are not measured on an interval scale, they must first be z-transformed prior to averaging using this formula (cf. Bortz 2005, 219):

$$z = \frac{1}{2}ln\left(\frac{1+r}{1-r}\right) \text{ (Eq. 6)}$$

To account for the fact that each verb realizes a different number of argument patterns, the resulting z-values of each verb are weighted by the respective number of argument realization patterns to calculate average z-values, using this formula:

$$\bar{z} = \frac{\sum_{j=1}^{k}(v_j - 3)z_j}{\sum_{j=1}^{k}(v_j - 3)} \text{ (Eq. 7)}$$

where v_j is the number of argument realization patterns for the verb j, and Z_j is the z-value calculated for this verb using Eq.6.. As Silver/Dunlap (1987) show, it is appropriate to backtransform the z-values to r-values by solving Eq. 6 for r:

$$\bar{r} = \frac{e^{2\bar{Z}} - 1}{e^{2\bar{Z}} + 1} \text{ (Eq. 8)}$$

Standard deviations (*SD*) are then calculated using the standard formula:

$$SD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(r_i - \bar{r})^2} \text{ (Eq. 9)}$$

where n is the number of verbs that belong to the cluster and r_i is the bivariate correlation coefficient for two corpora for verb i.

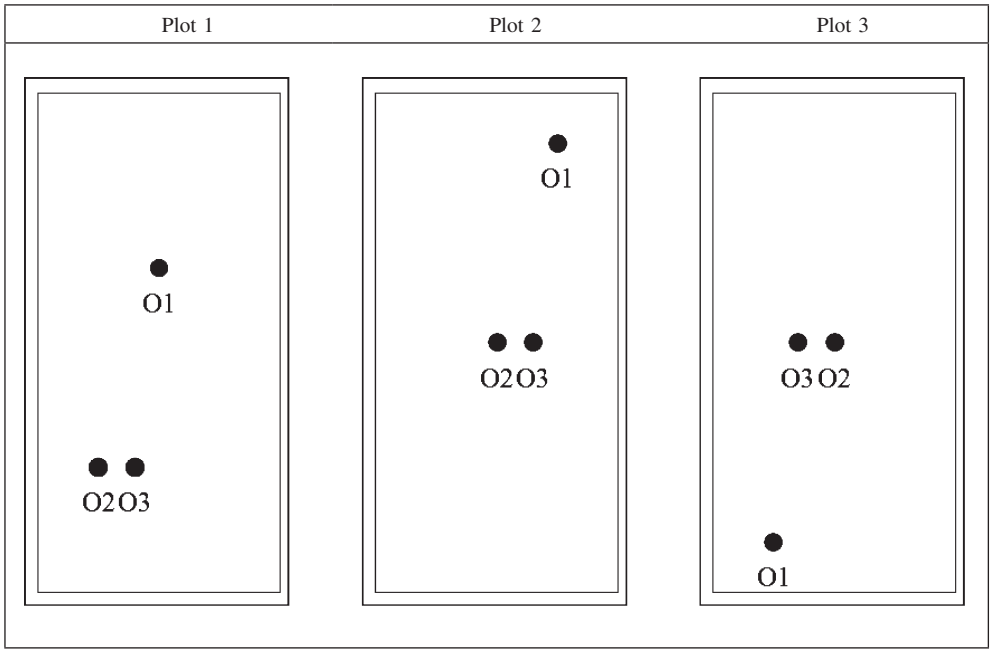


Figure 1: Example of three different configuration plots visualizing the same MDS solution.

This means that the axes of the plot are (in themselves) meaningless and that the orientation of the objects in the plot is arbitrary. So it is possible to perform any transformation (translation, reflection, or orthonormal rotation) of the configuration as long as this transformation does not affect the Euclidean distances (cf. StataCorp. 2011, 461). To illustrate this fact, Figure 1 shows three simple configuration plots that are identical in terms of an MDS solution.

If we interpret the result of the MDS, it does not matter which configuration plot we look at: in all three examples, object 2 and object 3 were mapped more closely to each other than either was to object 1.

To make the configuration plots obtained comparable, we transformed each of the resulting MDS configurations so that the coordinates of the Newspaper corpus are always fixed on [0.25,0.25] in the two-dimensional space. Additionally, it should be noted that the ratio of the width to height roughly represents the extent to which each of the two dimensions accounts for the dissimilarity found in the data (which is actually equivalent to a principal component analysis, cf. StataCorp. 2011, 443). In Figure 1 this means that the y-axis accounts for twice as much of the underlying distances as the x-axis. Furthermore, the proportion, that is the influence of the extracted principal components, is noted on the axes. For example, a value of $p = 75\%$ means that three-quarters of the underlying distances can be approximated by this dimension. So in the limit case ($p = 100\%$), this means that the complete underlying distances can be plotted on a straight line without losing any information.

4. Results

4.1 Results of the cluster analysis

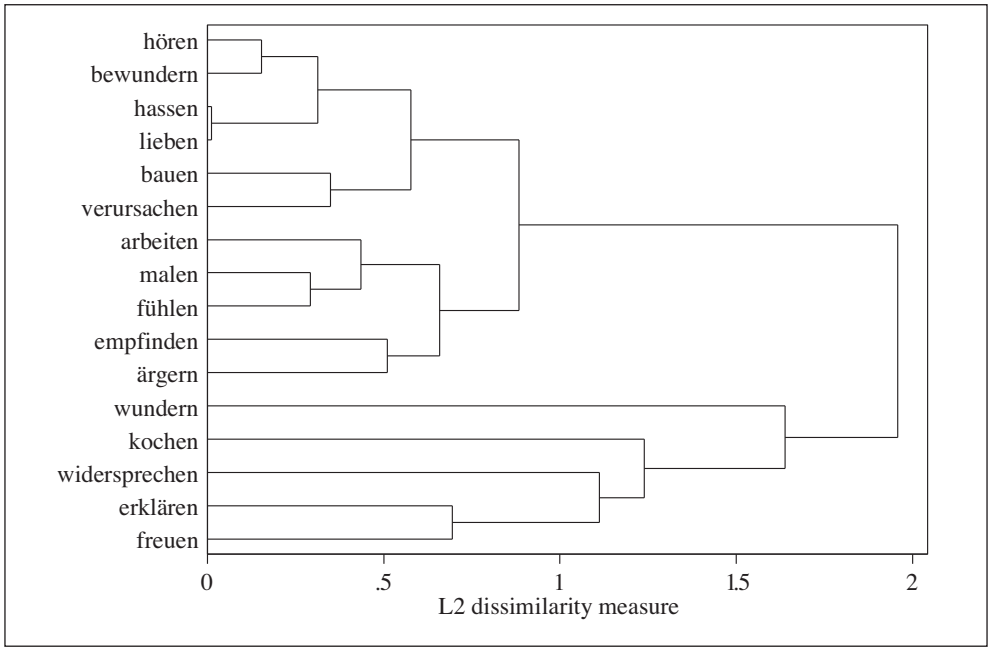
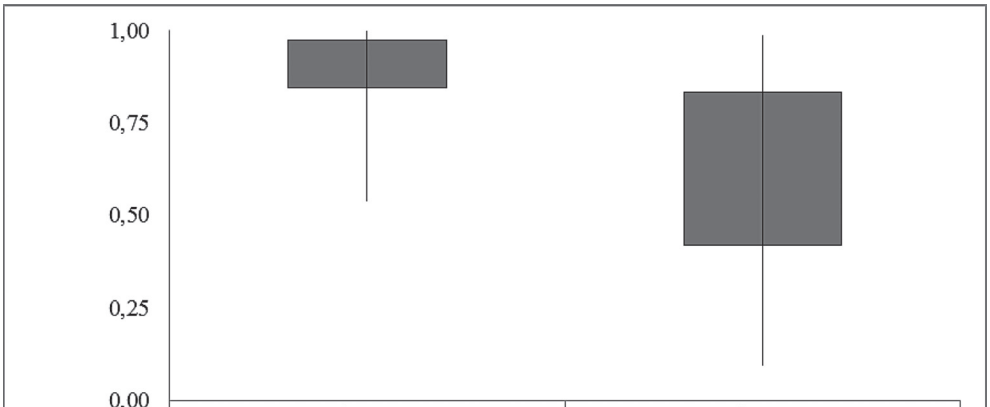


Figure 2: Dendrogram of the cluster analysis.

Figure 2 shows the results of the cluster analysis. The first result to be noted is that the dendrogram indicates the presence of two distinct clusters. The first bigger cluster (Cluster 1) comprises the verbs *hören* 'hear', *bewundern* 'admire', *lieben* 'love', *hassen* 'hate', *bauen* 'build', *verursachen* 'cause', *arbeiten* 'work', *malen* 'paint', *fühlen* 'feel', *empfinden* 'feel/sense' and *ärgern* 'get/make angry'. The rest of the verbs (*wundern* 'be astonished/astonish', *kochen* 'cook', *widersprechen* 'contradict', *erklären* 'explain' and *freuen* 'become/make happy') belong to the second cluster (Cluster 2). A closer data inspection shows that there is less variation for the verbs of Cluster 1 across corpora compared with Cluster 2. To visualize this result, we compared the following measures of dispersion between the two clusters in a box plot (cf. Figure 3): minimum, maximum, upper (75%) quartile, and lower (25%) quartile.

Since the difference between the upper and the lower quartile is equal to the middle 50% of the data values (the interquartile range), the different sizes of the boxes in Figure 3 show that there is less spread of correlation coefficients for verbs in Cluster 1 compared to Cluster 2. Furthermore, the boxes also show that the verbs in Cluster 1 are more similar regarding the distribution of argument realization patterns across corpora: 50% of all correlation coefficients for verbs in this cluster are within the range of 0.84 to 0.97. For Cluster 2, 50% of the correlation coefficients range from 0.42 to 0.83. The vertical bars



	Cluster 1	Cluster 2
Q75	0,97	0,83
Maximum	1,00	0,99
Minimum	0,54	0,10
Q25	0,84	0,42

Figure 3: Box plot for Cluster 1 and Cluster 2.

cutting the boxes show the maximum and the minimum for the cluster in question. Again this range is remarkably smaller for Cluster 1: while the smallest correlation coefficient for this cluster is $r = 0.54$, the minimum value for Cluster 2 is $r = 0.10$. Thus, the distributions of argument realization patterns for verbs in Cluster 1 are more similar in terms of correlation coefficients, and there is less cross-corpus variation for verbs in this cluster.

In order to further examine this relationship, the two clusters can be divided into smaller subclusters. Table 4 shows the result of this classification. Again this process shows that the verbs in Cluster 1 are more homogenous than the verbs in Cluster 2: while the first cluster comprises two subclusters (1.1, 1.2) with six and five verbs belonging to the respective clusters, the second cluster comprises four different clusters to which only one verb belongs (except for Cluster 2.1 which consists of two verbs). In other words, as the verbs of Cluster 2 are too different regarding the cross-corpus variation of the distribution of argument realization patterns, it makes no sense to cluster any of those verbs (again, except for Cluster 2.1) in a single common cluster.

Cluster	**Verbs**
1.1	*hören* 'hear', *bewundern* 'admire', *lieben* 'love', *hassen* 'hate', *bauen* 'build', *verursachen* 'cause'
1.2	*arbeiten* 'work', *malen* 'paint', *fühlen* 'feel', *empfinden* 'feel/sense', *ärgern* 'get/make angry'
2.1	*erklären* 'explain', *freuen* 'become/make happy'
2.2	*widersprechen* 'contradict'
2.3	*kochen* 'cook'
2.4	*wundern* 'be astonished/astonish'

Table 4: Subclusters and included verbs.

In the following section, we will present the results of MDS analyses for each of the six clusters to visualize the distribution of argument realization patterns and to interpret these results linguistically.

4.2 Results of multi-dimensional scaling

4.2.1 Cluster 1.1

Table 5 presents the pairwise correlation coefficients as a measure of cross-corpus similarity regarding the distribution of argument realization patterns. On the average, there is a high degree of similarity between the distributions for all verbs (*hören* 'hear', *bewundern* 'admire', *lieben* 'love', *hassen* 'hate', *bauen* 'build', *verursachen* 'cause') belonging to this cluster ($\bar{r}$ = 0.94, *SD* = 0.06): the similarity matrix shows that there are no noteworthy differences between the investigated corpora.

The two fiction corpora show the strongest linear correlation ($\bar{r}$ = 0.98, SD = 0.03) while the weakest emerging correlation between the NEWSPAPER corpus and the NON-FICTION corpus ($\bar{r}$ = 0.92, SD = 0.08) demonstrates that there is almost no cross-corpus variation regarding argument realization patterns for verbs belonging to this cluster. Figure 4 shows the resulting MDS configuration plot.

	NEWSPAPER	FICTION1	SPOKEN	FICTION2	NON-FICTION	SCIENCE
NEWSPAPER	1.00					
FICTION1	0.96	1.00				
SPOKEN	0.95	0.96	1.00			
FICTION2	0.93	0.98	0.97	1.00		
NON-FICTION	0.92	0.97	0.96	0.96	1.00	
SCIENCE	0.93	0.95	0.96	0.95	0.96	1.00

Table 5: Correlation matrix for Cluster 1.1.

The *Mardia fit 1* (*Mf1*) as a measure of goodness-of-fit (Stata Corp. 2011: 496) indicates that roughly 75% of the underlying proximities (i.e., similarities or dissimilarities) can be visualized by a two-dimensional configuration of the data. This is a reasonable fit. The *stress* (*s*) value measures the difference between the data of the input matrix and the resulting output configuration, so that the lower the *stress* the better the fit. In this case, the *stress* (*s* = 0.15) is reasonable, but no more than that (cf. Backhaus et al. 2003, 630). For example, in Figure 4, compared to the spatial proximity of the SPOKEN corpus and the FICTION2 corpus, the FICTION1 corpus and the FICTION2 corpus seem to be closer, but the correlation of both corpus pairs is almost identical, as the input correlation matrix demonstrates (cf. Table 5). However, this is mainly due to the general strong correlation between all corpora for verbs belonging to this cluster. In other words, knowing which corpora are correlated is of little help in predicting the strength of the resulting coefficient. This explains why our model finds it difficult to find an undistorted spatial configuration of the proximity matrix. Regarding the distribution of argument realization patterns, this also means that the

choice of corpus type does not seem to be an important factor for verbs belonging to this cluster.

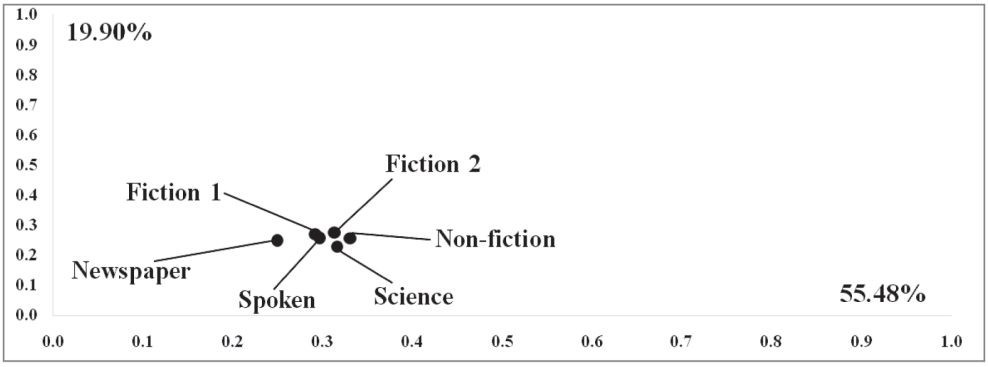


Figure 4: Two-dimensional configuration for Cluster 1.1 (classical MDS: s = 0.15; $Mf1$ = 0.75).

Apart from being similar with respect to their corpus correlations, the verbs in Cluster 1.1 also show very high correlations across the board. This is particularly striking with *lieben* and *hassen*, where all pairs of corpora have a correlation coefficient of at least 0.99. With respect to both verbs and all corpora, the simple NPNOM-NPACC-pattern accounts for at least 82% of the examples; the other sentences exhibit patterns with reflexive pronouns, complement sentences, PPs expressing a reason (2a), or passive constructions. *Bewundern* shows similar correlations (≥ 0,96), except for those corpus pairs involving the NEWSPAPER corpus (≤ 0,89), which differs from the other corpora in that it shows a high number of passive constructions (over 40% of the sample) (2b). Cross-corpus differences are also rarely found for *hören* (correlations between 0.89 and 0.98) though SPOKEN is somewhat different from the other corpora because it contains a higher number of simple transitive sentences (70% of the sample). The fiction corpora also show a slightly higher number of raising constructions (2c) for this verb.

(2) a. […] viele hassen ihn für seine brutalen Methoden.
many.NOM hate him.ACC for his brutal methods
'Many hate him for his brutal methods.'
[NEWSPAPER: B02/OKT.73537 Berliner Zeitung, 18.10.2002; Der Eiserne [3]]

b. Er wurde in Paris wegen seiner Begabung sehr bewundert […].
he.NOM PAST.PAS in Paris because of his talent very admired
'He was very much admired in Paris because of his talent.'
[NEWSPAPER: BVZ11/MAR.01407 Burgenländische Volkszeitung, 10.03.2011; Als Knabe in Paris bewundert]

c. Eine ganze Nacht lang hörte ich ihn einmal Poe und Byron englisch deklamieren.
A whole night long heard I.NOM him.ACC once Poe and Byron in English recite.
'I once heard him recite Poe and Byron in English all night long.'
[FICTION1: BIO/BKA.01895 Alfred Kerr: [Briefe 1895], In: Wo liegt Berlin? – Berlin, 1998 [36]]

As can be seen from the dendrogram (Figure 2), *verursachen* and *bauen* stand out from the other four verbs of the cluster. A closer look at the correlation tables reveals that these two verbs exhibit slightly larger differences between corpora than the four verbs mentioned above. With *verursachen*, NON-FICTION stands out because of its high proportion of passive sentences (40% of the sample) (3a) while the two fiction corpora exhibit a tendency to express an affected referent as a dative NP (3b). As with other verbs, *bauen* shows a preference for simple transitive sentences in spoken language while SCIENCE and NON-FICTION reveal a frequent use of passive constructions (3c).

(3) a. […] daß der Schaden von dem Gastwirt […] grob fahrlässig verursacht wird […].
that the damage.NOM by the landlord grossly negligently caused.PRES.PAS
'[…] that the damage is caused by the landlord through gross negligence.'
[NON-FICTION: Zimmermann, Theo, Der praktische Rechtsberater, Gütersloh: Bertelsmann 1957, 118]

b. […] der Leinölgeruch verursacht ihr Kopfschmerzen […].
the smell.NOM of linseed oil causes her.DAT headache
'The smell of linseed oil gives her a headache.'
[FICTION2: Strittmatter, Erwin, Der Laden, Berlin: Aufbau-Verl. 1983, 266]

c. Juristisch war sie zirkulär gebaut […].
legally be.PAST.PAS 3P.FEM.NOM circularly built
'Legally, it was built circularly.'
[NON-FICTION: Luhmann, Niklas, Die Gesellschaft der Gesellschaft, Frankfurt a.M.: Suhrkamp 1997, 27]

4.2.2 Cluster 1.2

Compared to Cluster 1.1, there is greater variation in the distribution of argument realization patterns for verbs that belong to Cluster 1.2, containing the verbs *arbeiten* 'work', *malen* 'paint', *fühlen* 'feel', *empfinden* 'feel/sense', and *ärgern* 'get/make angry' ($\bar{r}$ = 0.84, SD = 0.11; cf. Table 6). Again, the two corpora containing fictional texts show the greatest resemblance ($\bar{r}$ = 0.96, SD = 0.02) whereas the dissimilarity regarding the distribution of argument realization patterns between the SPOKEN corpus and the SCIENCE corpus is more pronounced, but still on a moderate level ($\bar{r}$ = 0.69, SD = 0.07).

	NEWSPAPER	FICTION1	SPOKEN	FICTION2	NON-FICTION	SCIENCE
NEWSPAPER	1.00					
FICTION1	0.91	1.00				
SPOKEN	0.85	0.89	1.00			
FICTION2	0.87	0.96	0.84	1.00		
NON-FICTION	0.91	0.93	0.88	0.91	1.00	
SCIENCE	0.80	0.78	0.69	0.76	0.86	1.00

Table 6: Correlation matrix for Cluster 1.2.

Figure 5 illustrates this relationship in an MDS configuration plot: on the one hand, the plot shows the noticeable difference between the two aforementioned corpora. On the other hand, the plot also shows similarity between all corpora except for the SCIENCE corpus.

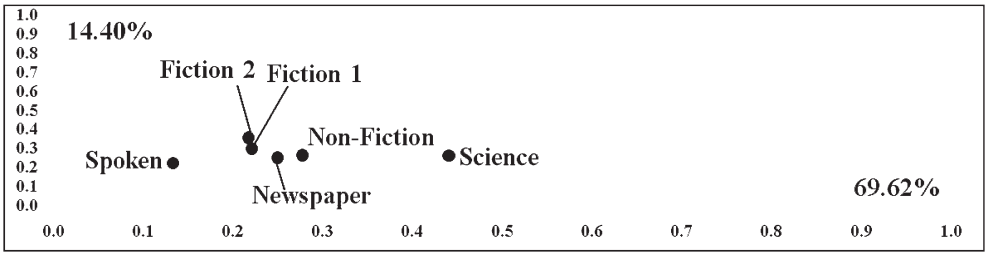


Figure 5: Two-dimensional configuration for Cluster 1.2 (classical MDS: *s* = 0.14; *Mf1* = 0.84).

The representation of distances in two dimensions accounts for 84% of the underlying distances. Though reasonable, the *stress* value is far from perfect (*s* = 0.13). Further data inspection reveals that the MDS extracted a third dimension explaining 8.27% of the input matrix. Therefore, an MDS solution with three dimensions better incorporates the correlation data (*s* = 0.05; *Mf1* = 0.92). However, this updated model does not alter the fact that most of the underlying dissimilarity (69.62%) can be approximated by the first dimension, visualizing mainly the contrast of the SPOKEN corpus and the SCIENCE corpus.

All verbs in Cluster 1.2 show a very high correlation between FICTION1 and FICTION2, the lowest correlation coefficient for all verbs being r=0.92[12]. All verbs in this cluster also show a high correlation between the two fiction corpora on the one hand and NEWSPAPER (lowest correlation coefficient: r=0.75) and NON-FICTION (lowest correlation coefficient: r= 0.77) on the other. With the exception of *empfinden*, all verbs in this cluster correlate strongly between both fiction corpora and SPOKEN, the minimum value for all verbs except *empfinden* being r=0.82. *Empfinden* shows a higher frequency of transitive patterns expressing the specification of an emotion (cf. 4a) in FICTION1 (30.0% of the sample) and FICTION2 (34.0% of the sample) than in the other corpora, while being used most frequently with Topic-Comment structures (cf. 4b) in the other corpora (35.5% on average for Topic-Comment constructions where the Comment is introduced by *als* ('as'); 8.9% on average for Topic-Comment structures where the Comment is realized as an adjectival phrase (cf. 4c)):

(4) a. Ich empfinde eigentlich mehr Scham als Angst, …
I.NOM sense actually more shame.ACC than fear
'I am actually more ashamed than afraid.'
[FICTION1: Victor Klemperer: [Tagebücher 1933], in: Ich will Zeugnis ablegen bis zum letzten, vol. 1 – Berlin, 1995 [15]]

b. Die Menschen empfinden die Katastrophe von Tschernobyl als großes Unglück.
The people.NOM sense the catastrophe.ACC of Chernobyl as big disaster.
'The people feel the catastrophe of Chernobyl to be a big disaster.'
[NON-FICTION: die tageszeitung, 15.12.1988, 10-11; Eine Reise in die "Zone"]

---

12 All data mentioned in the discussion of the different clusters may be obtained from AK (koplenig@ids-mannheim.de).

c. bloß das Sächsische empfinde ich im Vergleich zum Bayrischen […]
so total ungepflegt.
only the Saxonian.ACC sense I.NOM in comparison with Bavarian […]
so totally unrefined
'Only Saxonian do I feel to be so completely unrefined as compared to Bavarian.'
[SPOKEN: E:\IDS\KorporaGS\Dh_IV\NBB5_IV]

The verbs in Cluster 1.2 correlated least with respect to the distribution of their argument realization patterns between SCIENCE and SPOKEN, the mean correlation coefficient for these corpora being r_m=0.69. Of all verbs in Cluster 1.2, *arbeiten* shows the lowest correlation between SCIENCE and SPOKEN (r=0.54). For each of the verbs in Cluster 1.2, there are different reasons for the relatively low correlations between SCIENCE and SPOKEN:

arbeiten: While the agentive-intransitive use of *arbeiten* represents the most frequent pattern in all corpora, it is relatively rare in SCIENCE (28.4% of the sample). Active and passive patterns which do or do not contain an agent and express the subject worked on as a PP introduced by one of several prepositions, usually *über* (literally: over, here: 'on') or *zu* (literally: to, here: 'on'), are frequent in this corpus. Examples are *Sie arbeitet über das Thema* ('She is working on that topic') and *Über das Thema wurde viel gearbeitet* ('That topic has been worked on extensively'). Active and passive sentences of this type make up 55.2% of the SCIENCE sample, passive constructions accounting for 20.9% and active constructions for 34.3% of the sample. When the role of the subject worked on is realized as a PP headed by *an* ('at'), the sentence is interpreted as having a partitive interpretation, as in *Sie arbeiten an einem neuen Buch* ('They are working on a new book'). When patterns realizing the subject worked on by a PP headed by *an* are taken into account, in addition to those expressing the subject worked on as another type of PP, patterns expressing the subject worked on account for 64% of the SCIENCE sample gathered for *arbeiten*. If agentive-intransitive patterns are added to those expressing the employer, both patterns taken together account for 92.5% of the SCIENCE sample. The SPOKEN corpus is characterized by the frequency of sentences realizing the role of the employer as a PP headed by *in* ('in'), *bei* ('at'), *auf* ('on'), *für* ('for'), or *an* ('at'), as in *Sie arbeitet in einer Firma* ('She works in a company'). While these patterns account for 35.0% of the sample gathered from SPOKEN, they account for 11.9% on the average of the samples from the other corpora.

malen: The SCIENCE corpus shows only a few agentive-intransitive uses of *malen* as in *Sie malt* ('She is painting'/'She paints') (12.0% of the sample), but passive uses, as in *Etwas wird gemalt* ('Something is being painted'), are frequent in this corpus compared to other corpora (13.0% of the sample). By contrast, one-place agentive intransitive uses of *malen* are frequent in SPOKEN (42% of the sample).

empfinden: In SCIENCE, passive uses of the Topic-Comment structure expressing the Comment as a phrase introduced by *als* ('as') are frequent (18.9% of the sample for passive Topic-Comment structures without an experiencer, as in *Die Kirche wurde wieder als lebendiger Organismus empfunden* ['The church was again felt to be a living organism']; 7.1% for passive Topic-Comment structures expressing the experiencer as a PP headed by *von* ('by'), as in *Ihre Politik wurde von den Liberalen als vorbildlich empfunden* ['Her policy was felt by the Liberals to be exemplary']) while constructions expressing the specification of a feeling, as in *Man empfindet ein gewisses Unbehagen* ('One feels a certain uneasiness')

are relatively rare (7.1% of the sample). Typical of SPOKEN is the frequency of Topic-Comment structures expressing the Comment as an adjective, as in *Ich empfinde das eigentlich nur interessant* ('I actually only feel that to be interesting') (26.0% of the sample). In the other corpora, the Comment in Topic-Comment structures is mostly expressed as a phrase introduced by *als* ('as').

fühlen: The SCIENCE corpus shows a comparatively small proportion (30.7% of the sample) of reflexive patterns with an AP expressing the quality of the feeling experienced by the referent of the subject-NP, as in *Er fühlte sich müde* ('He felt tired') and a large proportion (29.7% of the sample) of transitive patterns with the accusative NP expressing the content of the experiencer's feeling as in *Er fühlte nur Mitleid* ('He felt only compassion'). SPOKEN shows exactly the opposite frequency pattern for these two argument realization patterns (73.0% and 1.0% of the sample, respectively).

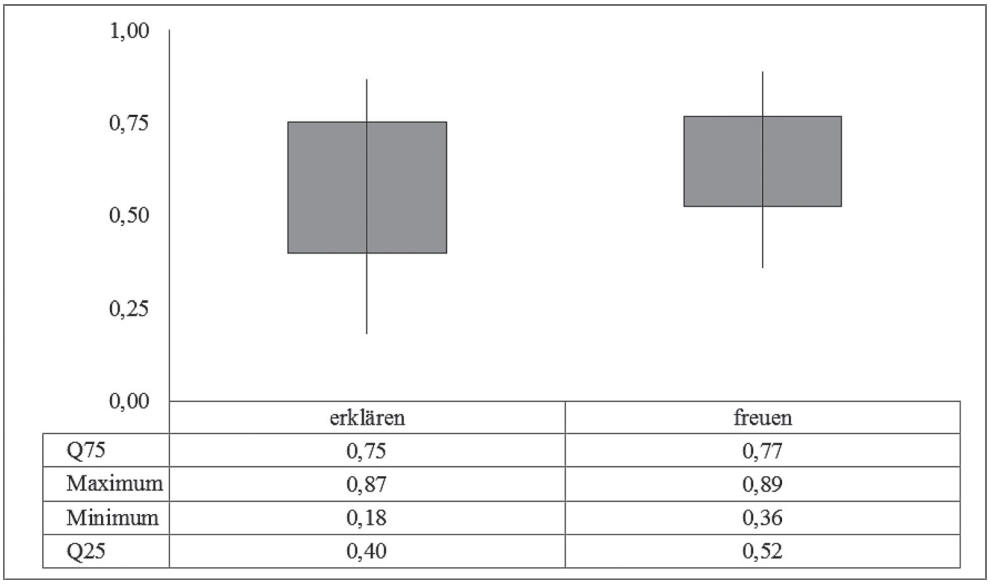
ärgern: SCIENCE shows a somewhat larger proportion (17.6% of the sample) of subordinate sentences introduced by *dass* ('that') which express the stimulus-argument (as in *Dass sie das nicht getan hat, ärgert ihn sehr* ('He is very angry about her not having done that') as well as a certain tendency towards patterns expressing an external experiencer, as in *Er ärgert sich über etwas* ('He is angry about something'). SPOKEN shows a slight preference for patterns expressing an external stimulus, as in *Das hat mich sehr geärgert* ('That made me very angry'). *Ärgern* is not used as a verb introducing direct speech either in SCIENCE or in SPOKEN.

On the whole, emotion verbs and perception verbs appear to be used primarily in their abstract senses, as in *etwas als etwas empfinden* ('sense something as something') and *Er fühlte nur Mitleid* ('He felt only compassion') in SCIENCE. Patterns expressing the specification of an emotion are rare in this corpus. Passive structures also appear to be used more readily in SCIENCE. One of the most salient characteristics of SPOKEN is a tendency towards the intransitive use of simple action verbs in their habitual senses. The frequency of these patterns is due to the fact that SPOKEN consists to a large extent of interviews where people are asked about their linguistic biographies. Many verbs which are commonly used to introduce direct speech in written language are not used in that way in spoken language. The verbs in Cluster 1.2 are only rarely used to introduce direct speech in SCIENCE and NON-FICTION.

4.2.3 Cluster 2.1

The two verbs *erklären* 'explain' and *freuen* 'become/make happy' form Cluster 2.1. A comparison of each verb's corpus correlation matrix documents the similarity: there are significant differences between the SPOKEN corpus and all the other corpora under examination. For instance, for both verbs, the minimum correlation is that between the SPOKEN corpus and the NEWSPAPER corpus ($r = 0.18$ for *erklären* 'explain' and $r = 0.36$ for *freuen* 'become/make happy'). At the same time, a comparison of the two correlation matrices also shows a difference of text genre: compared to the correlation between the SPOKEN corpus and the NON-FICTION corpus for the verb *erklären* 'explain', the correlation between the same two corpora is much stronger for *freuen* 'become/make happy' ($r = 0.44$ for *erklären*

'explain' and $r = 0.77$ for *freuen* 'become/make happy'). Apart from this difference, the cross-corpus variation of argument realization patterns is quite similar for both verbs (cf. Figure 6).



	erklären	freuen
Q75	0,75	0,77
Maximum	0,87	0,89
Minimum	0,18	0,36
Q25	0,40	0,52

Figure 6: Box plot for *erklären* 'explain', *freuen* 'become/make happy'.

The aggregated correlation matrix shows that, in addition to the deviation of the SPOKEN corpus, the SCIENCE corpus also stands out: except for NON-FICTION ($r = 0.67$), all other corpora correlate only weakly with this corpus (all rs < 0.55, cf. Table 7).

	NEWSPAPER	FICTION1	SPOKEN	FICTION2	NON-FICTION	SCIENCE
NEWSPAPER	1.00					
FICTION1	0.80	1.00				
SPOKEN	0.25	0.47	1.00			
FICTION2	0.76	0.83	0.45	1.00		
NON-FICTION	0.76	0.73	0.58	0.63	1.00	
SCIENCE	0.52	0.54	0.30	0.39	0.67	1.00

Table 7: Correlation matrix for Cluster 2.1.

The MDS configuration obtained fits the proximity data quite well ($s = 0.08$). It visualizes the input matrix in two dimensions accounting for roughly 87% of the underlying distances (cf. Figure 7).

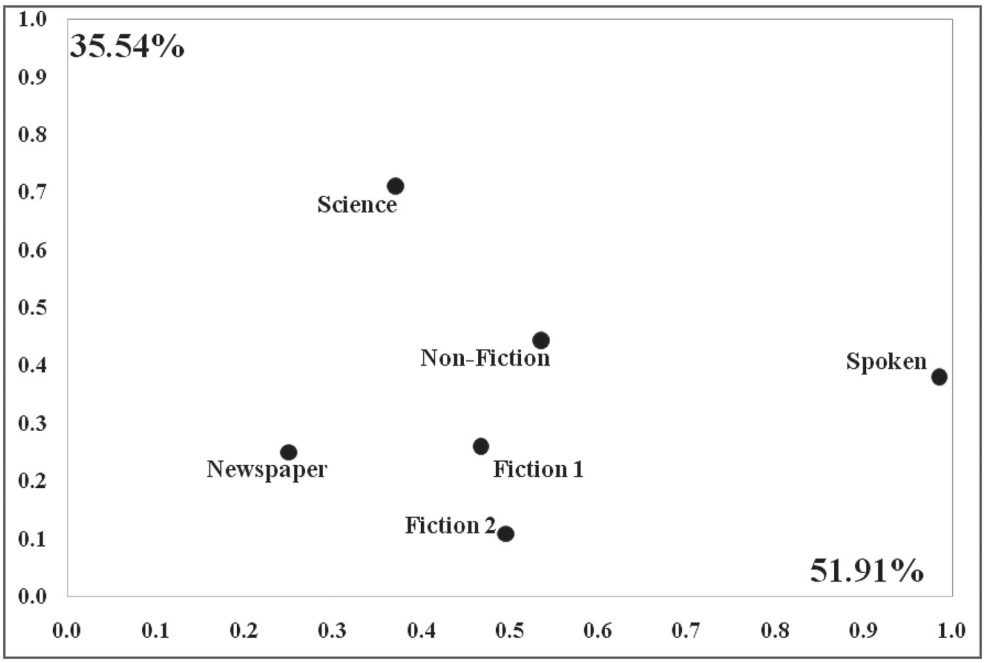


Figure 7: Two-dimensional configuration for Cluster 2.1 (classical MDS: s = 0.08; Mf1 = 0.87).

Both verbs show a low correlation between SPOKEN and all other corpora. However, different phenomena account for this distribution with respect to *freuen* on the one hand and *erklären* on the other. The pattern tables for spoken language exhibit a high proportion of simple transitive sentences expressing the agent and the explanandum with *erklären*, in particular when the addressee, which is expressed as a dative NP, is realized in addition to the agent and the explanandum (5a). Another important reason for the exceptional status of SPOKEN is that *erklären* does not occur as a verb introducing direct speech while examples of this sort make up more than 30% of the sample in NEWSPAPER and FICTION2. A third conspicuous feature of SPOKEN is the frequent occurrence of indirect interrogative clauses with *erklären* (5b).

(5) a. […] also ich erkläre denen das Berndeutsch […]
I.NOM explain them.DAT the Bern German.ACC
'[…] well, I explain "Bern" German to them […].'
[SPOKEN: E:\IDS\KorporaGS\Dh_IV\BER6_IV.TextGrid]

b. […] da wurde dann erklärt wie man richtig flirtet
there was then explained how to flirt properly
'[…] there it was explained then how to flirt properly.'
[SPOKEN: E:\IDS\KorporaGS\Dh_IV\NES1_IV.TextGrid]

Freuen belongs to those psych-verbs that show an alternation between a variant realizing the stimulus or the experiencer as subject. While in German the experiencer-as-subject variant seems to prevail generally with psych-verbs in newspaper texts (cf. Cosma/Engelberg

2013) (6a), spoken German shows a very low proportion of this variant with *freuen*. Furthermore, SPOKEN stands out with *freuen* because of the high proportion of patterns realizing the stimulus as a subordinate clause introduced by *wenn* ('when') (6b). *Freuen* is similar to *erklären* in that it does not occur as a verb introducing direct speech although it is used in that way in written language. Since *freuen* is not used particularly often in this function in general, this is only a minor factor.

(6) a. […] er] freut sich über die Unesco-Auszeichnung
he is happy REFL over the Unesco commendation
'He is happy about the Unesco commendation.'
[NEWSPAPER: T06/JAN.03640 die tageszeitung, 20.01.2006, 23; Unesco entdeckt "Designed in Berlin")]

b. […] die freut sich dann immer wenn sie Sächsisch reden kann
she is happy REFL then always when she Saxonian speak can
'then, she is always happy when she can speak Saxonian.'
[SPOKEN: E:\IDS\KorporaGS\Dh_IV\GLZ4_IV.TextGrid]

SCIENCE shows a rather low correlation with all the other corpora except with NON-FICTION. The reasons for this rather low correlation are different for *erklären* and *freuen*. *Erklären* shows a strikingly low proportion of patterns realizing the addressee as a dative NP. In contrast, the explanans is realized very frequently, in particular as a PP (7a). In contrast to NEWSPAPER, FICTION1, and FICTION2, *erklären* (as is probably also the case with other verbs) is only very rarely used in scientific discourse as a verb introducing direct speech. With *freuen*, in particular the rare use of the PP headed by *auf* (indicating the reading 'look forward to') accounts for the differences between SCIENCE and the other corpora (7b).

(7) a. Das Scheitern des Konkordats von 1817 erklärt sich aus der Tatsache, daß […].
the failure.NOM of the concordat of 1817 expalins itself out of the fact that
'The failure of the concordat of 1817 is explained by the fact that […].'
[SCIENCE: o.A., Die Kirche in der Gegenwart, Freiburg i. Br. [u.a.]: Herder 1971, 10555]

b. Die Männer freuten sich auf das Grillen […].
the men were happy REFL on the barbecue
'The men were looking forward to the barbecue […].'
[FICTION2: Jentzsch, Kerstin, Ankunft der Pandora, Berlin: Verl. Das Neue Berlin 1996, 138]

Finally, we observed that the two verbs differ with respect to their correlation between SPOKEN and NON-FICTION. The low correlation numbers for *erklären* are mainly due to the fact that NON-FICTION, like SCIENCE but in contrast to SPOKEN, contains a low number of examples realizing the addressee of *erklären* but a fairly high number of different types of complement clauses realizing the explanandum. On the other hand, the NON-FICTION sample for *freuen*, like SPOKEN but in contrast to SCIENCE, exhibits a frequent use of prospective *auf* (cf. 7b) and *wenn*-clauses realizing the stimulus (cf. 6b).

4.2.4 Cluster 2.2

Like Clusters 2.3 and 2.4, Cluster 2.2 comprises only one verb (*widersprechen* 'contradict'). While the NEWSPAPER corpus and the NON-FICTION corpus are almost identically distributed regarding the argument realization patterns ($r = 0.99$), the SCIENCE corpus and the FICTION2 corpus show the biggest differences in this context ($r = 0.23$, cf. Table 8). The MDS (s =

0.05; *Mf1* = 0.93) extracts one particularly influential dimension accounting for 83.37% of the dissimilarity (cf. Figure 8).

	NEWSPAPER	FICTION1	SPOKEN	FICTION2	NON-FICTION	SCIENCE
NEWSPAPER	1.00					
FICTION1	0.59	1.00				
SPOKEN	0.77	0.89	1.00			
FICTION2	0.43	0.75	0.58	1.00		
NON-FICTION	0.99	0.58	0.78	0.39	1.00	
SCIENCE	0.72	0.44	0.67	0.23	0.74	1.00

Table 8: Correlation matrix for Cluster 2.2.

One further feature stands out: compared to all other verbs investigated in this study, the bi-variate correlation between the two corpora containing fictional texts is the weakest for this verb (r = 0.75). Furthermore, all other corpora also correlate relatively weakly with the FICTION2 corpus (all rs < 0.58).

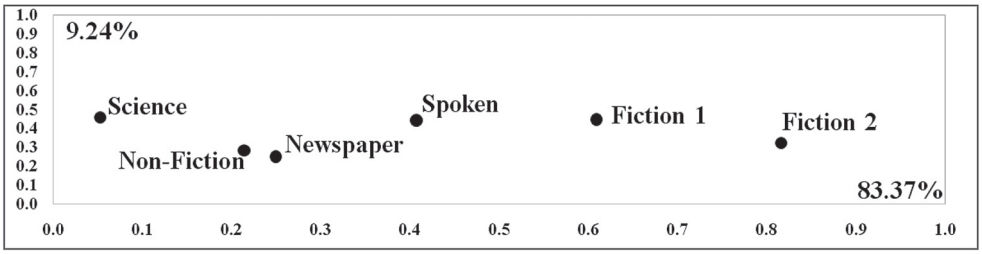


Figure 8: Two-dimensional configuration for Cluster 2.2 (classical MDS: s = 0.05; Mf1 = 0.93).

As is shown in Table 2, *widersprechen* is associated with two semantic arguments representing the contradicting propositions and two arguments representing the speaker and the addressee. The high correlation between NEWSPAPER and NON-FICTION (0.99) is due to the fact that in both corpora, a dative NP often expresses a propositional entity while a subject NP either refers to the speaker (about 30% of the sample) (8a) or to the contradicting proposition (about 30%) (8b).

(8) a. Andere „revisionistische“ Autoren widersprechen dieser Darstellung.
other “revisionist” authors.NOM contradict this account.DAT
‘Other “revisionist” authors contradict this account.’
[NEWSPAPER: F95/510.00010 Frankfurter Allgemeine, 1995]

b. Diese Bestimmung widerspricht nicht dem Art. 34 GG […].
this regulation.NOM contradicts not the article.DAT 34 GG
‘This regulation does not contradict article 34 GG […].’
[NON-FICTION: Zimmermann, Theo, Der praktische Rechtsberater, Gütersloh: Bertelsmann 1957, 186]

In contrast, FICTION2, which does not show a strong correlation with any of the other corpora, is characterized by low frequencies of the patterns represented by (8a) and (8b). On

the other hand, agentive uses without realization of the dative NP occur very frequently in FICTION2 (more than 60% of the sample) (9a), in half of the cases introducing direct speech (9b). FICTION1 shows a similar tendency, albeit weaker, towards dative-less agentive patterns.

(9) a. Es nützte nichts. Balla widersprach heftig.
Balla.NOM contradicted fiercely
'It didn't help. Balla contradicted him/her/them fiercely.'
[FICTION2: Neutsch, Erik, Spur der Steine, Halle (Saale): Mitteldeutscher Verl. 1964, 461]

b. "Quatsch", widersprach sie.
nonsense contradicted she.NOM
'"Nonsense", she contradicted.'
[FICTION2: Jentzsch, Kerstin, Ankunft der Pandora, Berlin: Verl. Das Neue Berlin 1996, 244]

The frequency of dative-less agentive patterns is one of the factors contributing to the particularly low correlation between SCIENCE and FICTION2. The latter is also due to the extremely low proportion of examples realizing agents as well as addressees with *widersprechen* in SCIENCE. At the same time, the bivalent pattern with non-human referents in nominative and dative position (as illustrated in 8b) occurs very often in scientific texts (67% of the sample).

Of all clusters and verbs, *widersprechen* exhibits the weakest correlation between the two fiction corpora. While both corpora show a strong tendency towards agentive uses with *widersprechen* (around 80%), the addressee dative occurs in 40% of the examples in FICTION1 but only in 18% of the sentences in FICTION2. In addition, *widersprechen* is used more often to introduce direct speech in FICTION2 (29% versus 10% in FICTION1). Both phenomena are related since the addressee is hardly ever realized when the verb is used to introduce direct speech. The difference between the two corpora is due to the composition of the corpora: in contrast to FICTION2, FICTION1 also includes autobiographical texts, in which direct speech probably occurs less often than in novels.

4.2.5 Cluster 2.3

Accounting for roughly 94% of the underlying distances, the MDS (s = 0.06, cf. Figure 9) carried out for Cluster 2.3, that is, the verb *kochen* 'cook', illustrates a verb idiosyncrasy: on the one hand, there is a relatively strong correlation both between the SCIENCE corpus and the NON-FICTION corpus (r = 0.84) and between the remaining corpora (all rs > 0.86). On the other hand, these two groups of corpora are clearly distinct from each other regarding cross-corpus similarity in the distribution of argument realization patterns (all rs < 0.54; cf. Table 9).

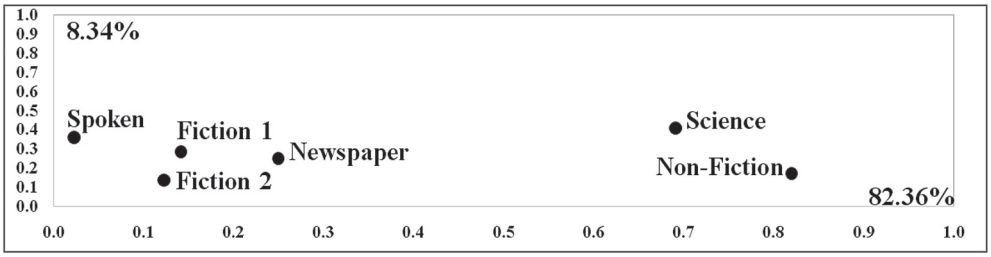


Figure 9: Two-dimensional configuration for Cluster 2.3 (classical MDS: s = 0.06; Mf1 = 0.91).

	Newspaper	Fiction1	Spoken	Fiction2	Non-fiction	Science
Newspaper	1.00					
Fiction1	0.91	1.00				
Spoken	0.86	0.90	1.00			
Fiction2	0.87	0.91	0.89	1.00		
Non-fiction	0.46	0.32	0.18	0.33	1.00	
Science	0.54	0.45	0.37	0.37	0.84	1.00

Tabelle 9: Correlation matrix for Cluster 2.3.

The difference observed is due to several related phenomema. Firstly, *kochen* exhibits a causative alternation yielding two different intransitive constructions. One of these expresses the person cooking[13] (10a, 10b) while the other realizes the thing or liquid boiling in subject position (10c).

(10) a. Kochen Sie Rhabarber nie in Aluminiumtöpfen.
cook you rhubarb never in aluminium pans
'Don't ever cook rhubarb in aluminium pans.'
[Newspaper: A97/APR.00582 St. Galler Tagblatt, 25.04.1997, Ressort: TB-LBN (Abk.); Der eigenwillige Sommerbote]

b. Du brauchst nicht zu kochen.
you need not to cook
'You don't have to cook.'
[Fiction2: Strittmatter, Erwin, Der Laden, Berlin: Aufbau-Verl. 1983, 266]

c. […] auch die Linsen werden gekocht […].
also the lentils.nom are being cooked
'[…] the lentils are being cooked, too […].'
[Non-fiction: Kölling, Alfred, Fachbuch für Kellner, Leipzig: Fachbuchverl. VEB 1956, 267]

In Science and Non-fiction the patient construction (mostly in passive constructions) exemplified by (10c) is preferred (ø 25% of the examples in the samples versus ø 2% in the other four corpora). Intransitive patient constructions in the active voice (such as *Das Wasser kocht* 'The water is boiling') have some predominance in the four other corpora (ø 7.3% versus ø 3% in science and Non-fiction). Newspaper, Fiction1, Fiction2, and Spoken show a preference for the agentive intransitive illustrated by (10b) (ø 38.2% versus ø 11.6% in the other two corpora). The same partition in the two sets of corpora can be observed if, in addition to the intransitive patterns realizing only the agent (10b) or only the patient (10c), intransitive patterns realizing additional semantic roles are also taken into account. Examples include (11a) and (11b).

(11) a. […] die kocht dann für uns alle.
she.nom cooks then for us all
'[…] then she will cook for us all.'
[spoken: Dh_IV\KUS2_IV.TextGrid]

---

13 This construction often has to be understood metaphorically in the sense of 'being very upset'.

b. Viele Puddingsorten können mit käuflichem Puddingpulver schnell gekocht werden.
many kinds.NOM of custard can with buyable custard powder fast be cooked
'Many kinds of custard can be made quickly with shop-bought custard powder…'
[NON-FICTION: Wir kochen gut, Leipzig: Verl für die Frau 1963, 176]

If all constructions are taken into account, NON-FICTION and SCIENCE show an average of ø 51.6% agentive sentences while the other corpora realize agents in as many as ø 84.9% of the sample sentences. The object being cooked is realized in ø 55.2% of the sentences in SCIENCE and NON-FICTION (including the argument expressing the resulting dish ø 71.4%) but only in ø 22.9% of the sentences in the other four corpora (including the resulting dish ø 51.6%). Apart from the distinction between agentive and non-agentive patterns, *kochen* also shows differences regarding the use of passive constructions in the different corpora: ø 41,9% of the sentences in SCIENCE and NON-FICTION versus ø 6,6 % of the sentences in the other corpora. Finally, beneficiary phrases co-occurring with *kochen* are distributed unevenly in the different corpora. The beneficiary is realized quite often in SPOKEN and in both FICTION corpora (between 5% and 13%) but is only moderately frequent (between 1% and 3%) in NEWSPAPER, SCIENCE, and NON-FICTION.

4.2.6 Cluster 2.4

Though different from all other verbs under examination, the cross-corpus variation for the verb *wundern* 'be astonished/astonish' belonging to Cluster 2.4 is in itself quite consistent (cf. Table 10): while there is a relatively strong inter-correlation for all corpora with written content (all *r*s > 0.74), the correlation between each of those corpora and the SPOKEN corpus is remarkably weak (all *r*s < 0.33).

	NEWSPAPER	FICTION1	SPOKEN	FICTION2	NON-FICTION	SCIENCE
NEWSPAPER	1.00					
FICTION1	0.87	1.00				
SPOKEN	0.21	0.33	1.00			
FICTION2	0.80	0.88	0.11	1.00		
NON-FICTION	0.88	0.85	0.14	0.74	1.00	
SCIENCE	0.86	0.88	0.10	0.86	0.91	1.00

Table 10: Correlation matrix for Cluster 2.4.

Consequently, the MDS (s = 0.07; *Mf1* = 0.94) retains one dimension that accounts for almost 90% of the proximity data and visualizes the specific characteristics for this verb in terms of the cross-corpus variation mentioned above (cf. Figure 10).

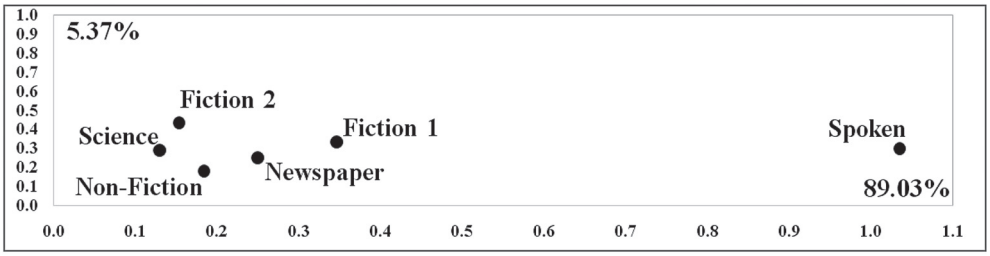


Figure 10: Two-dimensional configuration for Cluster 2.4 (classical MDS: s = 0.07; Mf1 = 0.94).

The low correlation between SPOKEN and the other corpora is due to the frequent use of patterns realizing the experiencer as an NP in the nominative case and the stimulus either as a subordinate clause introduced by *dass* ('that'), as in *Er wundert sich*, *dass …* ('He is surprised that …'), or as a PP headed by *über* (literally: on, over, here: 'about'), as in *Er wundert sich über ihre Frage* ('He is surprised about her question'), in written language (ø 24.37% and ø 21.63%, respectively, on the average of all the samples of written language). Both patterns occur only rarely in spoken language (1.54% of the SPOKEN sample for both patterns). Characteristic of spoken language is the frequent use of the pattern expressing the stimulus as an NP in the nominative case and the experiencer as an NP in the accusative case as in *Das hat ihn gewundert* ('That surprised him') (53.85% of the SPOKEN sample).

5. Discussion

5.1 Predictability of the distribution of patterns across corpora

This study started with (i) the observation that the distribution of the realization patterns of argument structures sometimes differs across corpora and (ii) the assumption that the homogeneity or inhomogeneity of cross-corpus behavior might depend on the semantic class a verb belongs to. This assumption could only marginally be confirmed. The class that behaved most homogeneously was that of directed emotion verbs (*bewundern* 'admire', *hassen* 'hate', *lieben* 'love') which shows hardly any cross-corpus differences. Action verbs (*arbeiten* 'work', *bauen* 'build', *kochen* 'cook/boil', *malen* 'paint') turned out to cluster quite closely, with the exception of *kochen*, which is the only one in the group that allows a causative alternation. Perception verbs (*fühlen* 'feel', *hören* 'hear', *empfinden* 'feel/sense') show slight similarities while neither connective verbs (*erklären* 'explain', *verursachen* 'cause', *widersprechen* 'contradict') nor alternating psych-verbs (*ärgern* 'get/make angry', *freuen* 'become/make happy', *wundern* 'be astonished/astonish') appeared to make up a cluster. Admittedly, a more fine-grained classification of verbs might have yielded a better correspondence between verb classes and the correlation of argument realization patterns between corpora.

However, even if the expected dependency of cross-corpus correlations on verb classes could not be fully shown, another parameter, namely, the particular type of argument

realization pattern realized, appeared to be decisive for the predictability of high or low correlations between corpora. Some types of realization pattern show a strong tendency towards low cross-corpora correlations, namely, (i) patterns realizing the agent as the subject when other patterns with non-agentive subjects are also available, (ii) patterns containing a dative NP referring to a human participant (particularly dative NPs realizing the role of an addressee, less so dative NPs realizing the role of a beneficiary), and (iii) patterns where the verb is used to introduce direct speech:

ad i) All of the five verbs in the second main branch of the dendrogram (see Figure 2, above), that is, in those subclusters showing on average the lowest corpus correlations and the largest span between the lowest and the highest pairwise corpus correlation, namely, Clusters 2.1 (r between 0.25 and 0.83), 2.2 (r between 0.23 and 0.99), 2.3 (r between 0.18 and 0.91), and 2.4 (r between 0.10 and 0.91) (*erklären*, *freuen*, *kochen*, *widersprechen*, *wundern*), exhibit an alternation between agentive and non-agentive subjects. Only two of the 11 verbs in the other subclusters (*ärgern*, *verursachen*) show an alternation of this kind.

ad ii) Both verbs allowing addressee datives (*widersprechen*, *erklären*) belong to the five verbs in the second main branch (Clusters 2.1 – 2.4).

ad iii) Finally, four of the five verbs that can introduce direct speech (*erklären*, *freuen*, *widersprechen*, *wundern*) belong to the second main branch. Only *ärgern*, which does not in any case frequently make use of this option, is a member of the first main branch.

In summary, this allows for the following prediction: whenever a verb shows an agent/non-agent alternation in subject position, allows for an animate dative-NP, or can be used to introduce direct speech, the distribution of argument realization patterns can be expected to differ widely across different text genres. The following section will discuss the extent to which this reflects the stylistic preferences of particular genres.

5.2 Genre-specific preferences for the realization of argument structure

We are not claiming that each of the corpora we compared represents a particular text genre. That would be a simplification of the intricacies of genre research. However, it is obvious that the corpora do not differ in an accidental way but show at least affinities with different text genres. By choosing two corpora that we considered to be of the same type, we were able to substantiate the assumption that text genre does indeed play a crucial role in the distribution of argument realization patterns. The corpora FICTION1 and FICTION2 proved to be more similar to each other than any other pair of corpora. In four of the five subclusters, of the 15 corpus pairs, FICTION1 and FICTION2 showed the strongest correlation. Even where FICTION1 and FICTION2 only exhibit a moderate correlation, namely, with Cluster 2.2, the difference seems to be the result of a difference in genre since it was due mainly to the presence or absence of autobiographical texts in FICTION1 and FICTION2, respectively.

SPOKEN: Besides SCIENCE, SPOKEN is the corpus which deviates the most with respect to the distribution of argument structure patterns. Firstly, it diverges from the other corpora in that it contains only very few instances of verbs used to introduce direct speech (0.3% of the sample as compared to 7.9% on average for the other corpora, cf. Table 11).

	Newspaper	Fiction1/2	Spoken	Non-fiction	Science
Direct Speech	16.89%	10.07%	0.33%	2.27%	2.29%

Table 11: Average proportion of direct speech introduction with the five verbs showing this pattern (*freuen* 'become/make happy', *wundern* 'be astonished/astonish', *ärgern* 'get/make angry', *widersprechen* 'contradict', *erklären* 'explain').

Secondly, spoken differs from the other corpora regarding the frequency of patterns expressing an addressee dative (46.4% of the sample versus ø 14.1% on average for the other corpora, cf. Table 12).

	Newspaper	Fiction1/2	Spoken	Non-fiction	Science
Addressee Dative	10.26%	28.87%	46.43%	10.49%	6.84%

Table 12: Average proportion of addressee datives[14] with the two verbs showing this pattern (*widersprechen* 'contradict' and *erklären* 'explain').

Thirdly, spoken shows a considerably larger proportion of simple transitive patterns for some of the verbs considered. Simple transitive patterns are frequent in spoken with *erklären* ('explain') and *bauen* ('build'), and especially with *wundern* ('be astonished/astonish') and *hören* ('hear'). The predominance of simple transitive and ditransitive patterns in Spoken has already been discussed with respect to *erklären* (see section 4.2). Simple transitive patterns are by far the most frequent pattern occurring with *bauen* in all corpora, but they are most frequent in Spoken, which contains 69.8% of all occurrences of these patterns, and in Newspaper, which comprises 62.0% of them. The proportion of simple transitive structures in Spoken is particularly high with *wundern* and *hören*: 53.8% of all transitive structures occurring with *wundern* and 69.3% of all transitive structures occurring with *hören* were found in Spoken, the next highest proportions being 10.0% and 45.0% (both in Fiction1), respectively. However, simple transitive patterns do not always predominate in Spoken. In the samples gathered for *fühlen*, for example, transitive patterns are most frequent in Science (29.7% of all occurrences of this pattern), Fiction2 (22.2% of all occurrences), and Fiction1 (15% of all occurrences), and are almost absent from Spoken and Newspaper (1.0% of all occurrences in both corpora). The frequency of transitive patterns occurring with *fühlen* in Science is likely to be due to the large proportion of philosophical texts in this corpus. Typical examples of transitive patterns in Science are *Man fühlt die Nähe Humes* ('One feels the proximity of Hume') and *Das Herz fühlt Gott* ('The heart feels God'). Examples of transitive patterns in Fiction1 and Fiction2 include *Sie fühlte seine Blicke* ('She felt his glances') and *Man fühlte die Absicht und die ganze Ferne des Vollbringens* ('One felt the intention and the whole distance of accomplishment'), which are typical of prose. While simple transitive patterns are almost absent from Spoken and Newspaper, the predominant pattern in these corpora is the pattern realizing the experiencer and the quality of the feeling, as in *Er fühlt sich einsam* ('He is feeling lonely').

---

[14] Without the autobiographical texts in Fiction1, the proportion of examples with addressee datives would probably be slightly lower.

Science: Besides the SPOKEN corpus, it is the SCIENCE corpus which shows the most striking differences to the other corpora. This concerns a number of argument structure features, most of which have already been mentioned in previous sections. Scientific texts contain a particularly high number of passive constructions (cf. Table 13).[15]

	NEWSPAPER	FICTION1/2	SPOKEN	NON-FICTION	SCIENCE
werden-PASSIVE	5.58%	4.84%	6.23%	14.68%	17.45%

Table 13: Average proportion of *werden*-passives with the 14 verbs exhibiting this pattern (*ärgern* 'get/make angry', *widersprechen* 'contradict', *erklären* 'explain', *verursachen* 'cause', *lieben* 'love', *hassen* 'hate', *bewundern* 'admire', *empfinden* ['feel/sense'], *fühlen* 'feel', *hören* 'hear', *arbeiten* 'work', *bauen* 'build', *kochen* 'cook', *malen* 'paint').

On the other hand, verbs in SCIENCE are rarely used to introduce direct speech (cf. Table 11). Agentive subjects occur significantly less often than in other corpora, in particular when used in simple intransitive sentences (cf. Table 14). Addressee datives are also rarely found in scientific texts (cf. Table 12).

	NEWSPAPER	FICTION1/2	SPOKEN	NON-FICTION	SCIENCE
INTRANSITIVE AGENT	29.01%	28.69%	30.89%	21.67%	11.75%

Table 14: Average proportion of intransitve agentive sentences realizing no further argument roles with the five verbs showing this pattern (*widersprechen* 'contradict', *arbeiten* 'work', *bauen* 'build', *kochen* 'cook', *malen* 'paint').

Fiction1 and Fiction2: The two fiction corpora are very homogenous with respect to text genres. Except for some autobiographical texts in FICTION1, they mainly consist of novels. Across the board, the fiction corpora did not prove to be particularly exceptional. The most characteristic trait of fiction texts seems to be that they pattern with NEWSPAPER and SPOKEN with respect to different argument structure features. As in spoken language, but in contrast to newspapers, fiction texts contain a high number of addressee datives – although not quite as many as in spoken language (cf. Table 12). As with newspaper texts, but in contrast to spoken language, they contain a high number of verbs introducing direct speech – although not quite as many as in newspaper texts (cf. Table 11). In contrast to SCIENCE and NON-FICTION, fictional texts pattern with newspaper texts and spoken language with respect to a high number of intransitive agentive uses (cf. Table 14) and a low proportion of passive structures (cf. Table 13).

---

15 These results differ from an early investigation on the frequency of the passive in German. Using a corpus of 15,000 sentences and a computer system based on punchcards, Brinker (1971) observed the following proportions of non-stative passive sentences (*werden*-passive) in texts of different genres: newspaper texts 9%, fiction 1.5%, general non-fiction 10.5%, scientific texts 9.4%; i.e., in contrast to our investigation, the passive in this corpus, which contains all different kinds of verbs, is less prominent in scientific texts compared to newspapers and other non-fiction texts. Building on Brinker's study, Schoenthal (1976) investigated the frequency of passive sentences in spoken language ("Freiburg corpus"): in 5.9% of the sentences, the *werden*-passive was used, which corresponds to the results of our study. Furthermore, Schoenthal observed a difference between public communication (6%) and private communication (3.3%).

Newspaper and Non-Fiction: Newspaper and Non-fiction show neither an especially low nor an especially high correlation either with the other corpora or with each other. Compared to each other, Non-fiction patterns a little more with Science than Newspaper does in showing a stronger tendency to the use of the passive (cf. Table 13) and a weaker affinity with direct speech (cf. Table 11). The two corpora are the most heterogeneous ones with respect to text genre. They contain a wide variety of different text genres. This is probably the reason for the unobtrusive correlation coefficients when the cross-corpus similarity is measured with regard to the distribution of argument realization patterns.

5.3 Impact on lexicological studies and lexicographical practice

In dictionaries, frequency information is quite often provided. Words or particular uses of words are characterized with labels such as *most*(*ly*), *frequent*(*ly*), *occasional*(*ly*), etc. (cf. Schaeder 1989), or explicit frequency information is given, as is often the case in recent online dictionaries. Our study has shown that with respect to valency and argument structure, the frequency of an item is often dependent on the genre of the text it occurs in. Interestingly, valency lexicography has rarely enriched the encoded valency frames with frequency information or information on text genres.[16] That might be due to the fact that – as we have seen – both are intricately connected and have to be extracted from corpora in a time-consuming manner. However, in encoding the often striking tendencies of language use, the two-dimensional combination of frequency and genre could add an interesting feature to descriptive lexicology.

Valency dictionaries are often considered to be of use in second language acquisition, in particular with respect to language production. As we have seen, the usage frequencies of the valency patterns of a verb differ widely. Employing data of this kind, valency dictionaries could not only order the valency frames they describe according to their frequencies, but they could also give explicit information about which frames are common and which are more exceptional. As our study has shown, with many verbs, this information would have to be relativized to text genres. Since language production even in a setting of second language learning rarely aims to produce context-free sentences but instead takes place as part of genre-specific tasks, information on what valency frame of a verb is common for a particular genre would lead to a properly contextualized acquisition of syntactic structures. Additional usage notes or specific outer texts of dictionaries which describe verb-unspecific preferences of text genres for particular valency patterns could add to this effect. Thus, from the point of view of pedagogical lexicography, which valency lexicography is often considered to be a part of, a connection of valencies to text genres and frequencies might allow for a more focused approach to the acquisition of syntactic patterns, as well as the specifics of text genres.

Apart from lexicography, the widely differing frequencies of argument realizations in different genres should of course also have an impact on theoretical approaches, in particular those that focus on the frequency-based entrenchment of linguistic entities in the linguistic system.

---

[16] VALBU (Schumacher et al. 2004) occasionally uses labels such as *häufig* ‘often’ and *selten* ‘rare’.

6. References

Backhaus et al. 2003 = Backhaus, Klaus / Erichson, Bernd / Plinke, Wolff / Weiber, Rolf: Multivariate Analysemethoden. Eine anwendungsorientierte Einführung. 10th ed., Berlin: Springer, 2003.

Barðdal 2008 = Barðdal, Jóhanna: Productivity. Evidence from Case and Argument Structure in Icelandic. Amsterdam, Philadelphia: Benjamins, 2008.

Behrens 2009 = Behrens, Heike: Konstruktionen im Spracherwerb. In: Zeitschrift für Germanistische Linguistik 37. 2009, 427–444.

Behrens 2011 = Behrens, Heike: Grammatik und Lexikon im Spracherwerb: Konstruktionsprozesse. In: Engelberg, Stefan / Holler, Anke / Proost, Kristel (edd.): Sprachliches Wissen zwischen Lexikon und Grammatik. Berlin, New York: de Gruyter, 2011, 375–396.

Boas 2010 = Boas, Hans C.: The syntax-lexicon continuum in Construction Grammar. A case study of English communication verbs. In: Belgian Journal of Linguistic 24. 2010, 54–82.

Boas 2011 = Boas, Hans C.: Coercion and leaking argument structures in Construction Grammar. In: Linguistics 49, 6. 2011, 1271–1303.

Bortz 2005 = Bortz, Jürgen: Statistik für Human- und Sozialwissenschaftler. 6th ed., Berlin: Springer, 2005.

Brinker 1971 = Brinker, Klaus: Das Passiv im heutigen Deutsch. Form und Funktion. Ismaning: Hueber, 1971.

Bybee 2010 = Bybee, Joan L.: Language, Usage and Cognition. Cambridge et al.: Cambridge University Press, 2010.

Bybee/Beckner 2010 = Bybee, Joan L. / Beckner, Clay: Usage-based theory. In: Heine, Bernd / Narrog, Heiko (edd.): The Oxford Handbook of Linguistic Analysis. Oxford: Oxford University Press, 2010, 827–855.

Carletta 1996 = Carletta, Jean: Assessing agreement on classification tasks: The kappa statistic. In: Computational Linguistics, 22, 2. 1996, 249–254.

Cosma/Engelberg 2013 = Cosma, Ruxandra / Engelberg, Stefan: Subjektsätze als alternative Valenzen im Deutschen und Rumänischen. In: Cosma, Ruxandra et al. (edd.): Komplexe Prädikationen als Argumente. Kontrastive Untersuchungen zum Deutschen, Rumänischen und Englischen. Berlin: Akademie-Verlag, 2013.

Engelberg 2009 = Engelberg, Stefan: Blätter knistern über den Beton. Zwischenbericht aus einer korpuslinguistischen Studie zur Bewegungsinterpretation bei Geräuschverben. In: Winkler, Edeltraud (ed.): Konstruktionelle Varianz bei Verben. OPAL, 4/2009. Mannheim: Institut für Deutsche Sprache. Online 2 June 2012: http://pub.ids-mannheim.de/laufend/opal/pdf/opal2009-4.pdf, 2009, 75–97.

Engelberg et al. 2011 = Engelberg, Stefan / König, Svenja / Proost, Kristel / Winkler, Edeltraud: Argumentstrukturmuster als Konstruktionen? Identität - Verwandtschaft - Idiosynkrasien. In: Engelberg, Stefan / Holler, Anke / Proost, Kristel (edd.): Sprachliches Wissen zwischen Lexikon und Grammatik. Berlin, New York: de Gruyter, 2011, 71–112.

Everitt et al. 2011 = Everitt, Brian S. / Landau, Sabine / Leese, Morven / Stahl, Daniel: Cluster Analysis. 5th ed., Chichester, UK: Wiley, 2011.

Gahl/Jurafsky/Roland 2004 = Gahl, Susanne / Jurafsky, Daniel / Roland, Douglas. Verb subcategorization frequencies: American English corpus data, methodological studies, and cross-corpus comparisons. In: Behavior Research Methods, Instruments, & Computers 36, 3. 2004, 432–443.

Geyken 2007 = Geyken, Alexander: The DWDS corpus: A reference corpus for the German language of the 20th century. In: Fellbaum, Christiane (ed.): Idioms and Collocations: Corpus-based Linguistic and Lexicographic Studies. London, New York: Continuum, 2007, 23–40.

Goschler 2011 = Goschler, Juliana: Geräuschverben mit direktionaler Erweiterung im Deutschen: Syntax, Semantik und Gebrauch. In: Lasch, Alexander / Ziem, Alexander (edd.): Konstruktionsgrammatik III. Aktuelle Fragen und Lösungsansätze. Tübingen: Stauffenburg, 2011, 27–41.

Gries 2006 = Gries, Stefan Th.: Exploring variability within and between corpora: some methodological considerations. In: Corpora 1, 2. 2007: 109–151.

Gries 2010 = Gries, Stefan Th.: Dispersions and adjusted frequencies in corpora: further explorations. In: Gries, Stefan Th. / Wullf, Stefanie / Davies, Mark (edd.): Corpus-linguistic Applications. Current Studies - New Directions. Amsterdam, New York: Rodopi, 2010 197–211.

Gries 2011 = Gries, Stefan Th.: Corpus data in usage-based linguistics. What's the right degree of granularity for the analysis of argument structure constructions? In: Brdar, Mario / Gries, Stefan Th. / Fuchs, Milena Žic (edd.): Cognitive Linguistics. Convergence and Expansion. Amsterdam, Philadelphia: Benjamins, 2011, 237–256.

Gries/Stefanowitsch 2004 = Gries, Stefan Th. / Stefanowitsch, Anatol: Extending collostructional analysis. A corpus-based perspective on 'alternations'. In: International Journal of Corpus Linguistics 9, 1. 2004, 97–129.

Gries/Stefanowitsch 2010 = Gries, Stefan Th. / Stefanowitsch, Anatol: Cluster Analysis and the Identification of Collexeme Classes. In: Newman, John / Rice, Sally (edd.): Empirical and Experimental Methods in Cognitive/Functional Research. Stanford: CSLI, 2010, 59–71.

Helbig 1985 = Helbig, Gerhard: Valenz und Kommunikation (ein Wort zur Diskussion). In: Deutsch als Fremdsprache, 22, 3. 1985, 153–156.

Jacobs 2009 = Jacobs, Joachim (2009): Valenzbindung oder Konstruktionsbindung? Eine Grundfrage der Grammatiktheorie. In: Zeitschrift für germanistische Linguistik 37, 3. 2009, 490–513.

Köhler 2005 = Köhler, Reinhard: Quantitative Untersuchungen zur Valenz deutscher Verben. In: Glottometrics 9, 3. 2005, 13–20.

Ludwig-Mayerhofer 2005 = Ludwig-Mayerhofer, Wolfgang: Korrelation und Assoziation. In: ILMES – Internet-Lexikon der Methoden der empirischen Sozialforschung. Online 30 April 2012: http://www.lrz-muenchen.de/~wlm/ein_voll.htm, 2005.

Maienborn 1996 = Maienborn, Claudia: Situation und Lokation: Die Bedeutung lokaler Adjunkte von Verbalprojektionen. Studien zur deutschen Grammatik; 53. Tübingen: Stauffenburg, 1996.

Roland 2001 = Roland, Douglas William: Verb Sense and Verb Subcategorization Probabilities. Doctoral dissertation, University of Colorado, 2001.

Roland/Jurafsky 1998 = Roland, Douglas / Jurafsky, Daniel: How verb subcategorization frequencies are affected by corpus choice. In: COLING-ACL '98: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Montreal, Quebec, Canada. Montreal, Aug. 10–14, 1998, 1122–1128.

Roland et al. 2000 = Roland, Douglas / Jurafsky, Daniel / Menn, Lise / Gahl, Susanne / Elder, Elizabeth / Riddoch, Chris: Verb Subcategorization Frequency Differences between Business-News and Balanced Corpora: The Role of Verb Sense. In: Comparing Corpora. A workshop held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics, Hongkong, 7th October 2000, 28–34.

Schaeder 1989 = Schaeder, Burkhard: Diafrequente Markierungen im allgemeinen einsprachigen Wörterbuch. In: Hausmann, Franz Josef / Reichmann, Oskar / Wiegand, Herbert E. (edd.): Wörterbücher. Dictionaries. Dictionnaires. Ein internationales Handbuch zur Lexikographie, Volume 1, Berlin, New York: de Gruyter, 1989, 688–693.

Schatte 2002 = Schatte, Christoph: Zur Theorie eines fachtextsortenspezifischen Valenzlexikons. In: Kovtyk, Bogdan / Wendt, Gabriele (edd.): Aktuelle Probleme der Übersetzungswissenschaft. Frankfurt/M.: Lang, 2002, 77–83.

Schlüter 2006 = Schlüter, Norbert: How Reliable are the Results? Comparing Corpus-Based Studies of the Present Perfect. In: Zeitschrift für Anglistik und Amerikanistik 54, 2. 2006, 135-148.

Schmid 2010 = Schmid, Hans-Jörg: Does frequency in text instantiate entrenchment in the cognitive system? In: Glynn, Dylan / Fischer, Kerstin (edd.): Quantitative Methods in Cognitive Semantics: Corpus-Driven Approaches. Berlin, New York: de Gruyter, 2010, 101–133.

Schoenthal 1976 = Schoenthal, Gisela: Das Passiv in der deutschen Standardsprache. Darstellung in der neueren Grammatiktheorie und Verwendung in Texten gesprochener Sprache. Ismaning: Hueber, 1976.

Schulte im Walde 2003 = Schulte im Walde, Sabine: Experiments on the Automatic Induction of German Semantic Verb Classes. Stuttgart: Universität Stuttgart: Institut für Maschinelle Sprachverarbeitung, 2003.

Schulte im Walde 2009 = Schulte im Walde, Sabine: The induction of verb frames and verb classes from corpora. In: Lüdeling, Anke / Kytö Merja Kytö (edd.): Corpus Linguistics. An International Handbook, Volume 2. Berlin, New York: Mouton de Gruyter, 2009, 952–971.

Schumacher et al. 2004 = Schumacher, Helmut / Kubczak, Jacqueline Kubczak / Schmidt, Renate / de Ruiter, Vera: VALBU - Valenzwörterbuch deutscher Verben. Tübingen: Narr, 2004.

Schwittalla 1985 = Schwittalla, Johannes: Verbvalenz und Text. In: Deutsch als Fremdsprache 22. 1985, 266–270.

Silver/Dunlap 1987 = Silver, N. Clayton / Dunlap, William P.: Averaging correlation coefficients: Should Fisher's z transformation be used? In: Journal of Applied Psychology, 72. 1987 146–148.

Sommerfeldt 1993 = Sommerfeldt, Karl-Ernst: Sprachliche Felder - Valenz - Textsorte. In: Wirkendes Wort, 43, 2. 1993, 317–336.

Sommerfeldt 1999 = Sommerfeldt, Karl-Ernst: Textsortenwandel und Valenz. In: Skibitzki, Bernd / Wotjak, Barbara (edd.): Linguistik und Deutsch als Fremdsprache. Festschrift für Gerhard Helbig zum 70. Geburtstag. Tübingen: Niemeyer, 1999, 189–200.

StataCorp 2011 = Stata Corp.: Stata Statistical Software: Release 12. College Station, TX: StataCorp LP, 2011.

Stefanowitsch 2011 = Stefanowitsch, Anatol (2011): Argument Structure: Item-Based or Distributed? In: Zeitschrift für Anglistik und Amerikanistik 59, 4. 2011, 369–386.

Stefanowitsch & Gries 2003 = Stefanowitsch, Anatol / Gries, Stefan Th.: Collostructions: Investigating the interaction of words and constructions. In: International Journal of Corpus Linguistics 8. 2003, 209–243.

Tomasello 2003 = Tomasello, Michael: Constructing a Language: A Usage-Based Theory of Language Acquisition. Cambridge, MA: Harvard University Press, 2003.

Tomasello 2006 = Tomasello, Michael: Acquiring Linguistic Constructions. In: Siegler, Robert / Kuhn, Deanna (edd.): Handbook of Child Psychology. Volume 2: Cognitive Development. New York: Wiley, 2006, 255–298.

Welke 2009 = Welke, Klaus: Valenz und Konstruktionsgrammatik. In: Zeitschrift für Germanistische Linguistik 37. 2009, 81–124.

Winkler 2009 = Winkler, Edeltraud (ed.): Konstruktionelle Varianz bei Verben. OPAL, 4/2009. Mannheim: Institut für Deutsche Sprache. Online 2 June 2012: http://pub.ids-mannheim.de/laufend/opal/pdf/opal2009-4.pdf, 2009.