Constructing a Valence Lexicon for a Treebank of German

Erhard W. Hinrichs, Kathrin Beck

{eh, kbeck}@sfs.uni-tuebingen.de

University of Tübingen
Seminar für Sprachwissenschaft
Germany

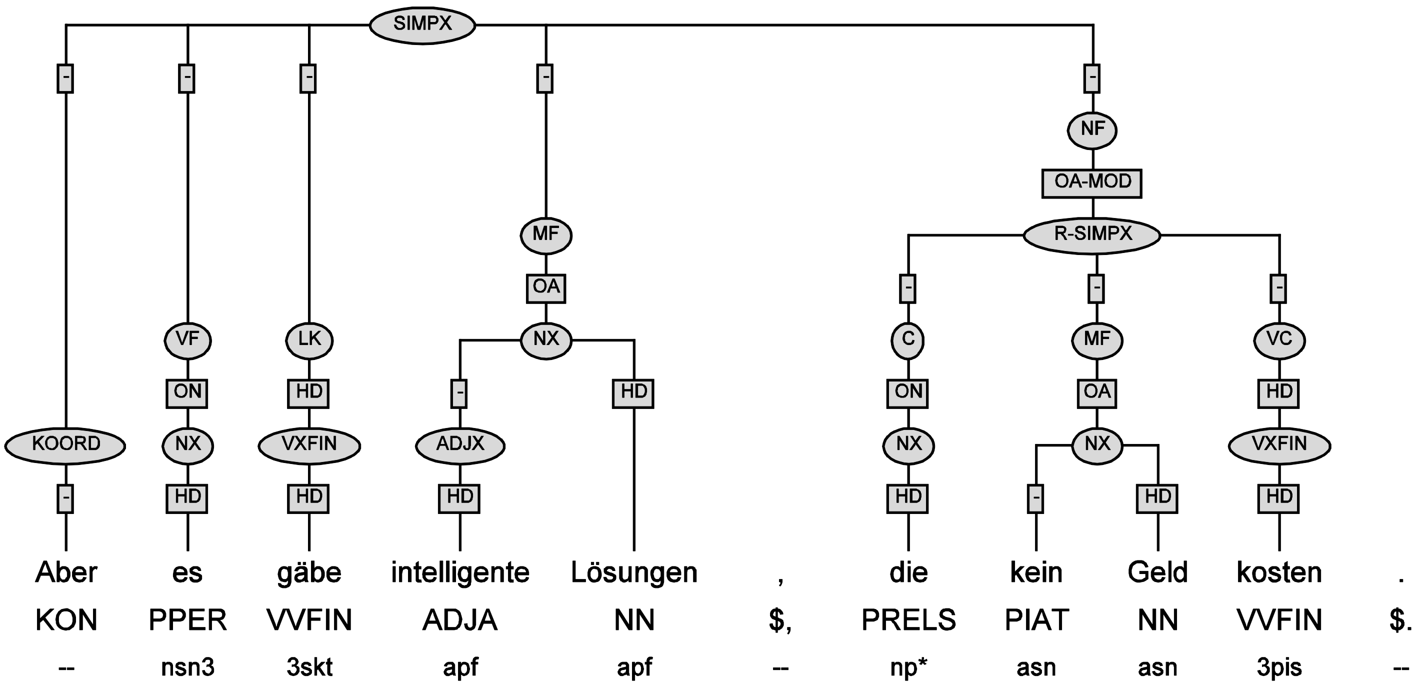
The TüBa-D/Z Treebank

German newspaper corpus:

- data source: die tageszeitung (taz)
- ca. 36 000 sentences
- semi-automatic annotation

Annotation scheme:

- context-free backbone
- PS grammar + predicate argument structure
- topological fields



‘But there would be intelligent solutions which do not cost money.’

Other Valence Lexica

- PropBank (Palmer et al. 2005)

  additional layer of semantic roles in the Penn Treebank

- FrameNet (Baker et al. 1998)

  based on frame semantics

- Prague valency lexicon PDT-VALLEX (Hajič et al. 2003)

  created on the basis of the Prague Dependency Treebank

The TüBa-D/Z Valence Lexicon

The valence lexicon:

- constructed in lockstep with the development of the TüBa-D/Z
- The number of verb lemmas and valence frames corresponds with the number of sentences in the TüBa-D/Z
- 4896 distinct verb lemmas
- 8013 valence frames (total)
- 717 distinct valence frames

Example entry of a polysemous verb:

einsetzen:
=======

ON [einsetzen] OA (R4-5603)
Bsp: Wir haben Computer eingesetzt
'We used the computer.'

ON [einsetzen] OA FOPP (für, gegen) (R4-3126)
Bsp: Wir setzen uns für eine Feuerpause ein
'We supported a cease fire.'
Bsp: Gegen den Widerstand setzt der Senat Polizeiknüppel ein (R4-27058)
'Against the resistance the senate used billy clubs.'

ON [einsetzen] (R4-2903)
Bsp: Schneefall hatte eingesetzt
'Snowfall had set in.'

ON [einsetzen] OA PRED (R4-17034)
Bsp: Gourmetköche setzen sie als Garnitur ein
'Gourmet cooks used it as garnish.'

ON [einsetzen] OD OA (N5-37382)
Bsp: Man setzt den Pflanzen neue Gene ein
'One inserts new genes into the plants.'

Grammatical Function Labels

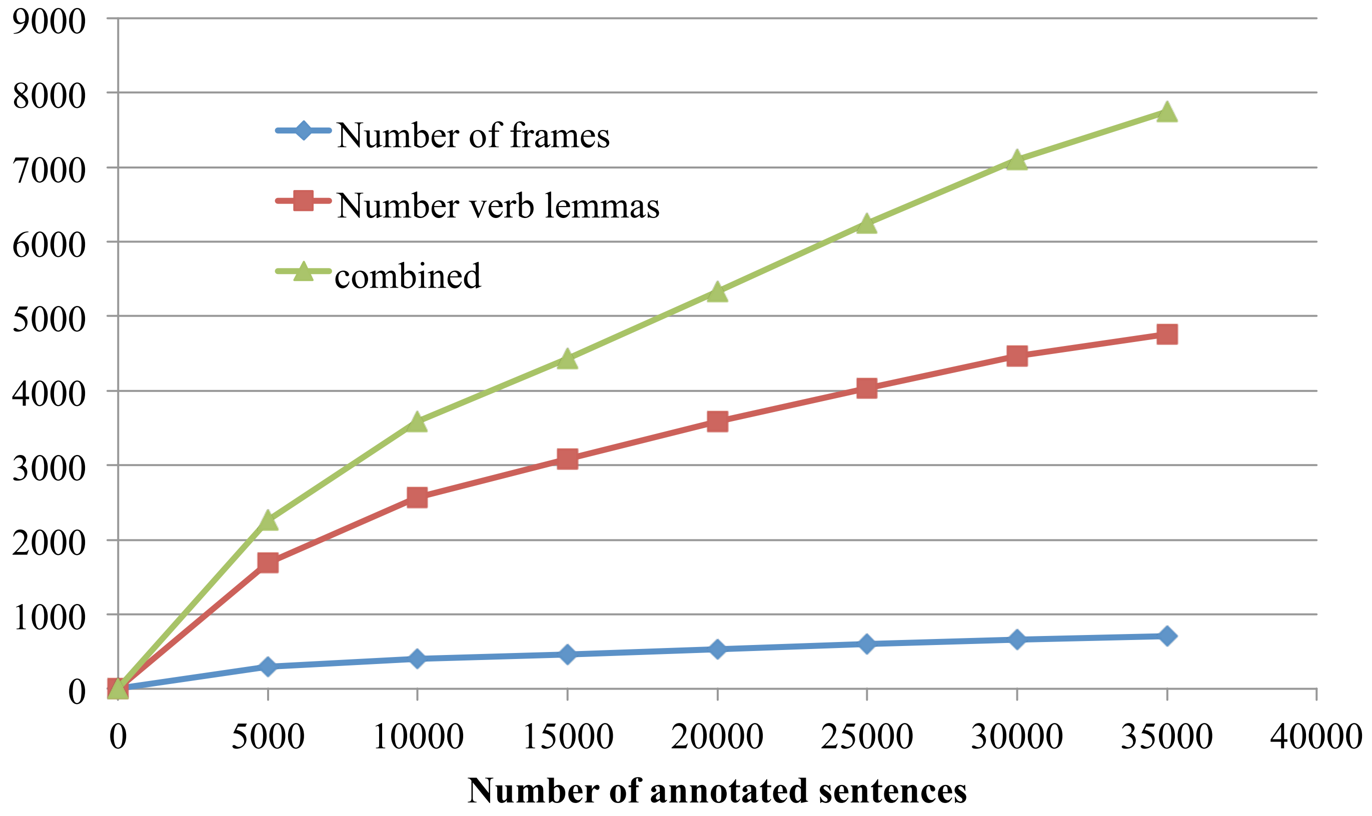
Inventory of grammatical function labels used in the valence lexicon:

- ➢ coincides with the edge labels used in the syntactic annotation
- ➢ corresponds directly to syntax

Label	**Description**
ON	nominative object (incl. subject clauses)
OG	genitive object
OD	dative object
OA	accusative object
OS	sentential object
OPP	obligatory prepositional object
FOPP	facultative prepositional object
OADVP	adverbial object
OADJP	adjectival object
PRED	predicate
OV	verbal object

Quantitative Analysis I

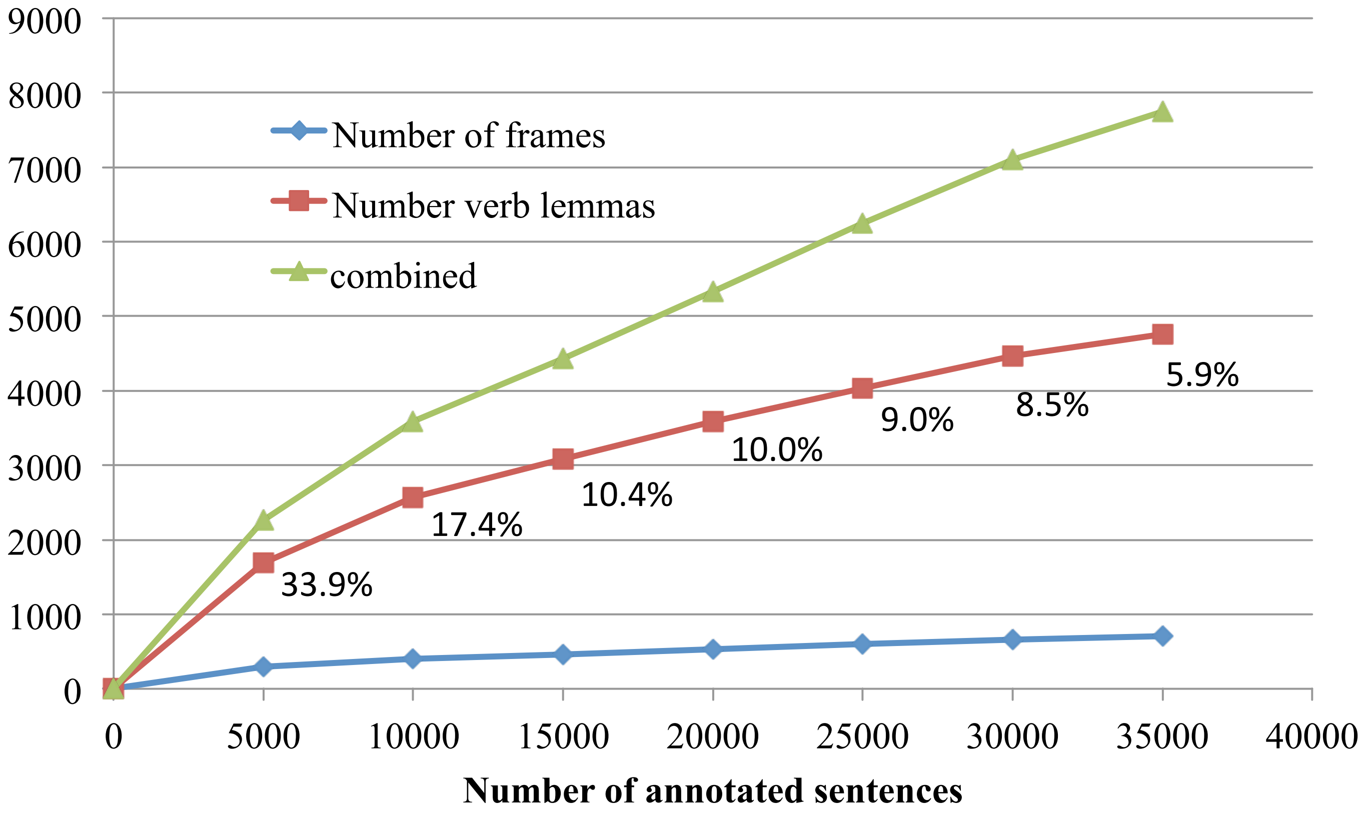
Accession rates for frames, verb lemmas, and their combinations in ranges of 5000 sentences:



Quantitative Analysis I

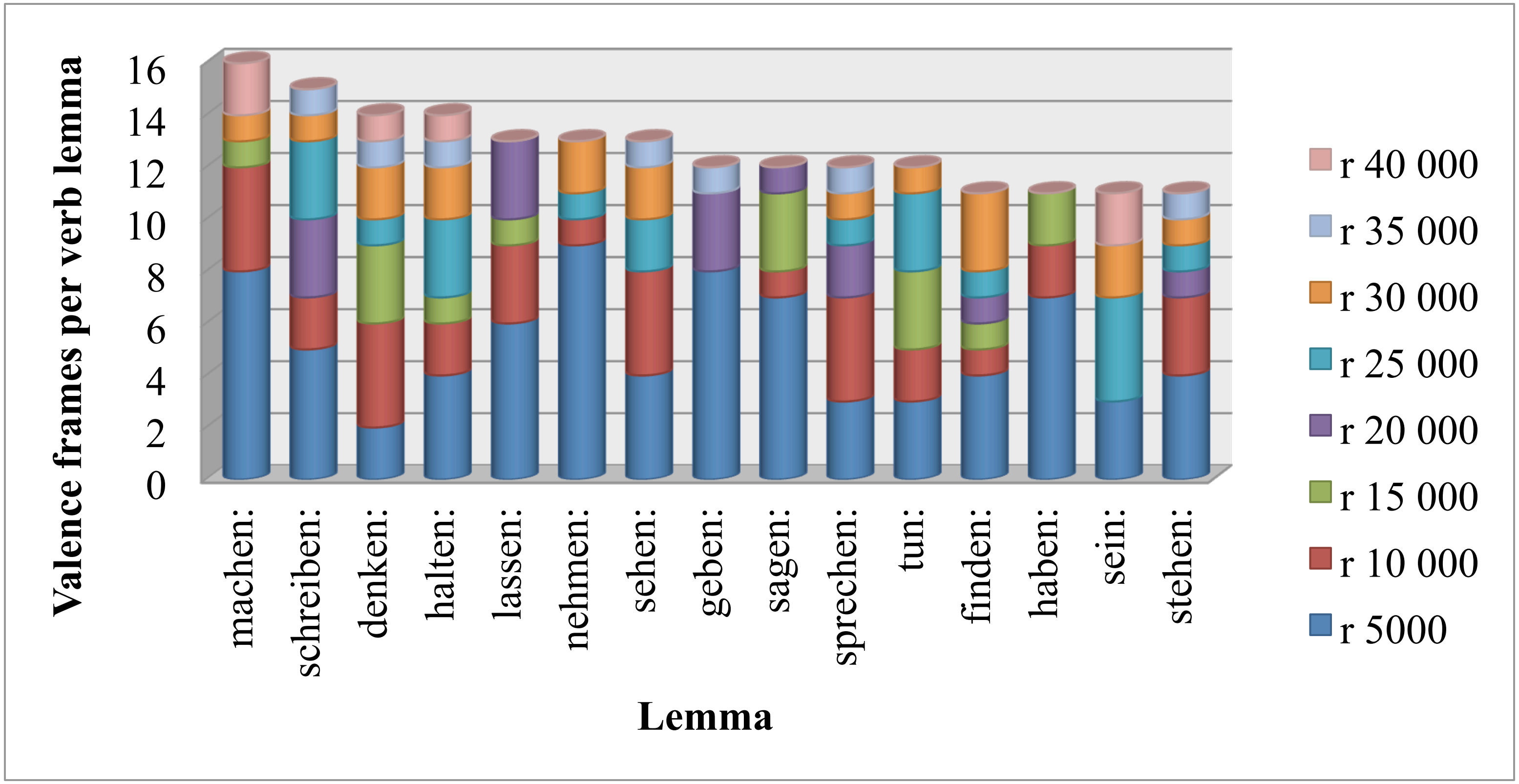
Eberhard Karls Universität Tübingen

Accession rates for frames, verb lemmas, and their combinations in ranges of 5000 sentences:



Quantitative Analysis II

Distribution of valence frames over sentence number range (r) for the 15 verb lemmas with the highest number of valence frames:



Quantitative Analysis III

Number of distinct valence frames:

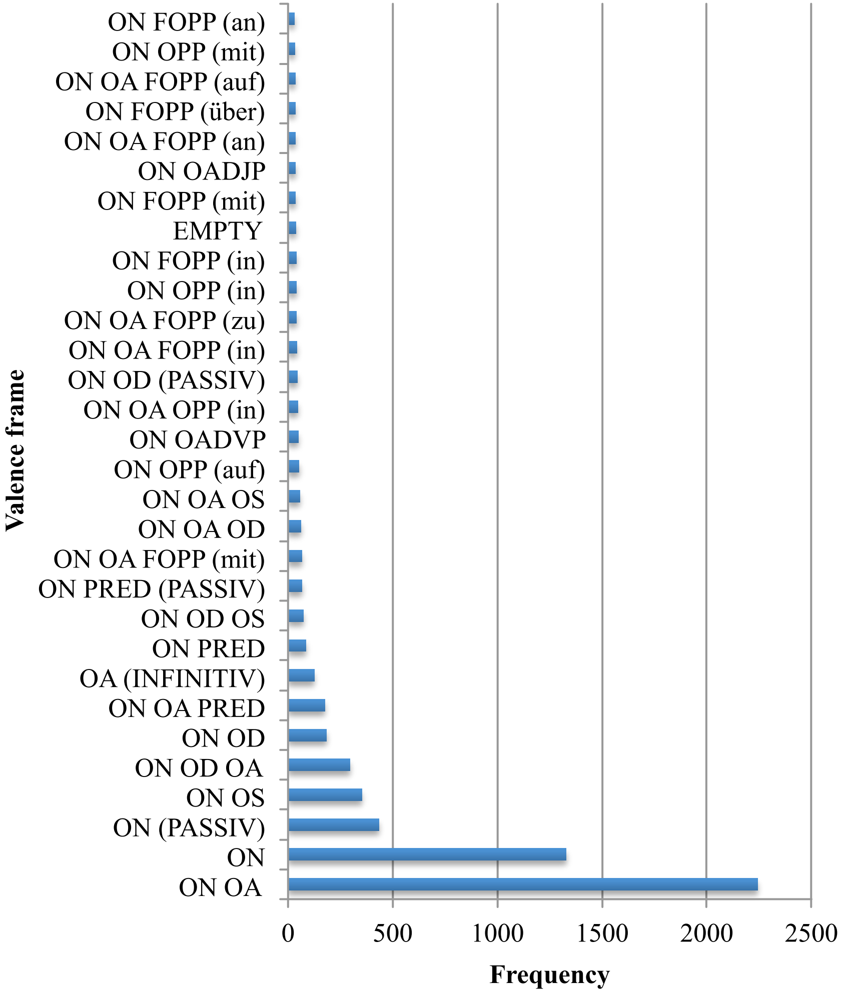
- 717 distinct valence frames (including prepositions)

- The frequency of occurrence for a specific valence frame ranges from

  2243 (ON OA)
  down to

  3 (36 distinct valence frames)
  2 (67 distinct valence frames)
  1 (488 distinct valence frames)

Top 30 list of valence frames



Quantitative Analysis IV

Valence frame count per verb lemma and frequency count:

4896 verb lemmas (total)

67.3%	(3294 verb lemmas):	1 frame
18.8%	(921 verb lemmas):	2 frames
7.1%	(347 verb lemmas):	3 frames
3.0%	(146 verb lemmas):	4 frames
1.7%	(85 verb lemmas):	5 frames
1.8%	(88 verb lemmas):	6-10 frames
0.3%	(15 verb lemmas):	more than 10 frames

Verb lemma	Valence frames per verb lemma	Frequency count
machen	16	1
schreiben	15	1
denken, halten	14	2
lassen, nehmen, sehen	13	3
geben, sagen, sprechen, tun	12	4
finden, haben, sein, stehen	11	4
entscheiden ... wissen	10	9
bleiben … verpflichten	9	6
bekommen … ziehen	8	15
anfangen … zahlen	7	25
abstimmen … wünschen	6	33
anbieten … zwingen	5	85
abfahren … zustimmen	4	146
abgeben … zweifeln	3	347
abbrechen … zutreffen	2	921
aalen … zwitschern	1	3294

Conclusion and Future Work

Current state of work:

➢ TüBa-D/Z:	ca. 40 000 sentences
➢ Valence Lexicon:	4947 distinct verb lemmas 8139 valence frames (total) 755 distinct valence frames

Integration with other resources of German (e.g. GermaNet):

Benefits:

➢ opportunity to clarify the intended sense of a verb by matches of verb senses with valence frames

➢ empirical verification of the relationship between the correlation of distinct valence frames and sense distinction

Thank you

for your attention

Quantitative Analysis V

Correlation of lemma frequency with the number of valence frames per verb:

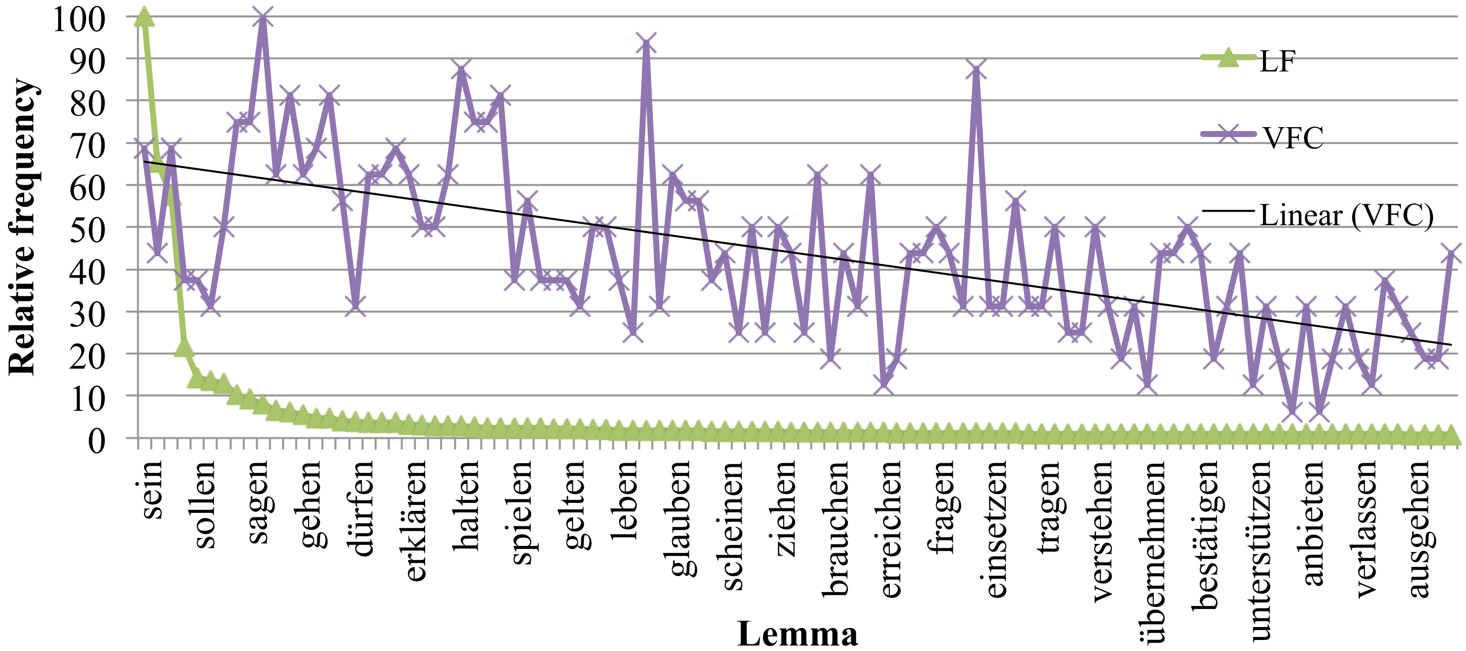
Top 20 correlation of lemma frequency and valence frame count per verb

Lemma	Lemma frequency	Valence frame count per verb
sein	10009	11
werden	6545	7
haben	5766	11
können	2164	6
sollen	1418	6
müssen	1373	5
wollen	1294	8
geben	1021	12
sagen	922	12
machen	801	16
kommen	668	10
lassen	626	13
gehen	562	10
stehen	475	11
sehen	462	13
bleiben	409	9
dürfen	379	5
heißen	364	10
wissen	364	10
finden	361	11

Quantitative Analysis VI

Top 100 correlation of lemma frequency and valence frame count:

- weak correlation


Relative frequency
LF
VFC
Linear (VFC)
sein
sollen
sagen
gehen
dürfen
erklären
halten
spielen
gelten
leben
glauben
scheinen
ziehen
brauchen
erreichen
fragen
einsetzen
tragen
verstehen
übernehmen
bestätigen
unterstützen
anbieten
verlassen
ausgehen
Lemma