empirikom

DFG Deutsche Forschungsgemeinschaft

Wissenschaftliches Netzwerk „Empirische Erforschung internetbasierter Kommunikation“

Gewinnung, Aufbereitung und Analyse von Korpora zu Genres internetbasierter Kommunikation: Herausforderungen und Perspektiven

technische universität dortmund

FAU FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

TECHNISCHE UNIVERSITÄT DARMSTADT

DIPF

Michael Beißwenger

Stefan Evert

Torsten Zesch

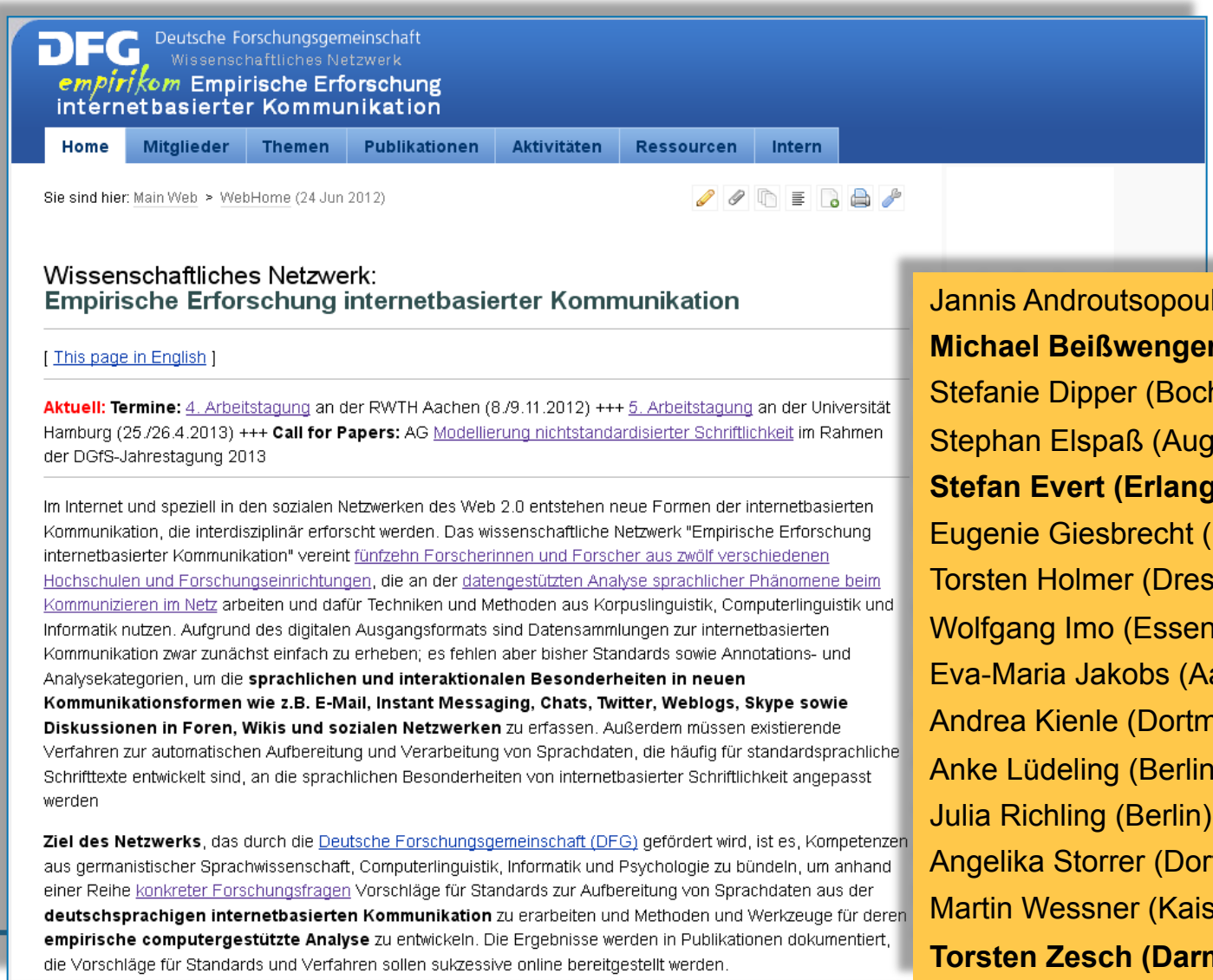
Workshop: Webkorpora in Computerlinguistik und Sprachforschung // Mannheim, 27./28. Sept. 2012

GSCL

IDS INSTITUT FÜR DEUTSCHE SPRACHE

Wissenschaftliches Netzwerk empirikom

DFG-Netzwerk Empirische Erforschung internetbasierter Kommunikation (2010-2013) http://www.empirikom.net



DFG Deutsche Forschungsgemeinschaft
Wissenschaftliches Netzwerk
empirikom Empirische Erforschung internetbasierter Kommunikation

Home | Mitglieder | Themen | Publikationen | Aktivitäten | Ressourcen | Intern

Sie sind hier: Main Web > WebHome (24 Jun 2012)

Wissenschaftliches Netzwerk:
Empirische Erforschung internetbasierter Kommunikation

[This page in English]

Aktuell: Termine: 4. Arbeitstagung an der RWTH Aachen (8./9.11.2012) +++ 5. Arbeitstagung an der Universität Hamburg (25./26.4.2013) +++ **Call for Papers:** AG Modellierung nichtstandardisierter Schriftlichkeit im Rahmen der DGfS-Jahrestagung 2013

Im Internet und speziell in den sozialen Netzwerken des Web 2.0 entstehen neue Formen der internetbasierten Kommunikation, die interdisziplinär erforscht werden. Das wissenschaftliche Netzwerk "Empirische Erforschung internetbasierter Kommunikation" vereint fünfzehn Forscherinnen und Forscher aus zwölf verschiedenen Hochschulen und Forschungseinrichtungen, die an der datengestützten Analyse sprachlicher Phänomene beim Kommunizieren im Netz arbeiten und dafür Techniken und Methoden aus Korpuslinguistik, Computerlinguistik und Informatik nutzen. Aufgrund des digitalen Ausgangsformats sind Datensammlungen zur internetbasierten Kommunikation zwar zunächst einfach zu erheben; es fehlen aber bisher Standards sowie Annotations- und Analysekategorien, um die **sprachlichen und interaktionalen Besonderheiten in neuen Kommunikationsformen wie z.B. E-Mail, Instant Messaging, Chats, Twitter, Weblogs, Skype sowie Diskussionen in Foren, Wikis und sozialen Netzwerken** zu erfassen. Außerdem müssen existierende Verfahren zur automatischen Aufbereitung und Verarbeitung von Sprachdaten, die häufig für standardsprachliche Schrifttexte entwickelt sind, an die sprachlichen Besonderheiten von internetbasierter Schriftlichkeit angepasst werden

Ziel des Netzwerks, das durch die Deutsche Forschungsgemeinschaft (DFG) gefördert wird, ist es, Kompetenzen aus germanistischer Sprachwissenschaft, Computerlinguistik, Informatik und Psychologie zu bündeln, um anhand einer Reihe konkreter Forschungsfragen Vorschläge für Standards zur Aufbereitung von Sprachdaten aus der **deutschsprachigen internetbasierten Kommunikation** zu erarbeiten und Methoden und Werkzeuge für deren **empirische computergestützte Analyse** zu entwickeln. Die Ergebnisse werden in Publikationen dokumentiert, die Vorschläge für Standards und Verfahren sollen sukzessive online bereitgestellt werden.

Jannis Androutsopoulos (Hamburg)
Michael Beißwenger (Dortmund)
Stefanie Dipper (Bochum)
Stephan Elspaß (Augsburg)
Stefan Evert (Erlangen)
Eugenie Giesbrecht (Karlsruhe)
Torsten Holmer (Dresden)
Wolfgang Imo (Essen)
Eva-Maria Jakobs (Aachen)
Andrea Kienle (Dortmund)
Anke Lüdeling (Berlin)
Julia Richling (Berlin)
Angelika Storrer (Dortmund)
Martin Wessner (Kaiserslautern)
Torsten Zesch (Darmstadt)

IBK-Korpora in Sprachforschung und Computerlinguistik: Perspektiven

Für welche Bereiche linguistischer und computerlinguistischer Forschung ist der Aufbau von IBK-Korpora und die Entwicklung von Verfahren zur Aufbereitung und Analyse von IBK-Daten relevant?

- Für die korpusgestützte Analyse sprachlicher Besonderheiten in der internetbasierten Kommunikation (IBK).
- Für die korpusgestützte Analyse aktueller Tendenzen in der deutschen Gegenwartssprache: Die Integration von IBK-Daten bzw. -Teilkorpora in annotierte Korpora zur deutschen Gegenwartssprache ermöglicht u.a. Untersuchungen zum Sprachwandel *durch* IBK.
- Für alle, die in Linguistik, Computerlinguistik und Informatik mit linguistisch aufbereiteten Webkorpora arbeiten (und dabei auch mit IBK-Phänomenen umgehen müssen).

IBK-Korpora in Sprachforschung & Computerlinguistik: Herausforderungen

Welche Fragen und Herausforderungen stellen sich beim Aufbau von IBK-Korpora und bei der linguistischen Aufbereitung von IBK-Daten?

- Fragen der Auswahl, Erhebung und Dokumentation von IBK-Daten
- Juristische und forschungsethische Fragen in Bezug auf die wissenschaftliche Nutzung von IBK-Daten und ihre Bereitstellung in Korpora
- Fragen der Modellierung und Repräsentation von IBK-Genres und -Dokumenttypen (→ s.a. *DeRiK*-Vortrag im Anschluss)
- Fragen geeigneter Analyse- und Beschreibungskategorien für IBK-spezifische Strukturmerkmale und Stilelemente
- Fragen der linguistischen Verarbeitung und Annotation (Tokenisierung, Satzgrenzenerkennung, Lemmatisierung, POS-Annotation, ...) (→ s.a. Vortrag Geyken/Jurish/Würzner)
- ...

Einblicke in die empirikom-Werkstatt

Zoom-in in drei Bereiche gegenwärtiger Forschung:

- POS-Tagging für internetbasierte Schriftlichkeit: Problemaufriss auf Basis von Experimenten mit Wikpedia-, Chat- und Twitter-Daten
- Gewinnung von IBK-Korpora / Webkorpora: Herausforderungen und Lösungsansätze
- Experimente zur linguistischen Evaluation von Webkorpora

Einblicke in die empirikom-Werkstatt

Zoom-in in drei Bereiche gegenwärtiger Forschung:

- **POS-Tagging für internetbasierte Schriftlichkeit:** Problemaufriss auf Basis von Experimenten mit Wikpedia-, Chat- und Twitter-Daten

Typische sprachliche Besonderheiten in internetbasierter Kommunikation:

- Schnellschreib-Phänomene (tolerierte Tippfehler)
- sprachliche Ökonomie: liberaler Umgang mit orthographischen Normen, die auf eine Verständnissicherung in der Distanzkommunikation hin optimiert sind (z.B. GKS, Interpunktion); Akronyme
- Orientierung am Duktus der gesprochenen Umgangssprache (Lexik, Syntax)
- „Verschriftete Umgangssprache“: Verschriftungen, die sich an der umgangssprachlichen Lautung anstatt am schriftlichen Standard orientieren
- Verwendung innovativer semiotischer und sprachlicher Formen, die sich in der IBK als Mittel zur emotionalen und evaluativen Kommentierung, zur Kohärenz-sicherung und zum spielerischen Rekurs auf Körperlichkeit herausgebildet haben (Emoticons, Inflektive, Adressierungsausdrücke)

POS-Tagging für IBK-Daten: Tests mit gängigen NLP-Tools

Testdatenset mit Belegen für ausgewählte Phänomene IBK-spezifischer Sprachverwendung

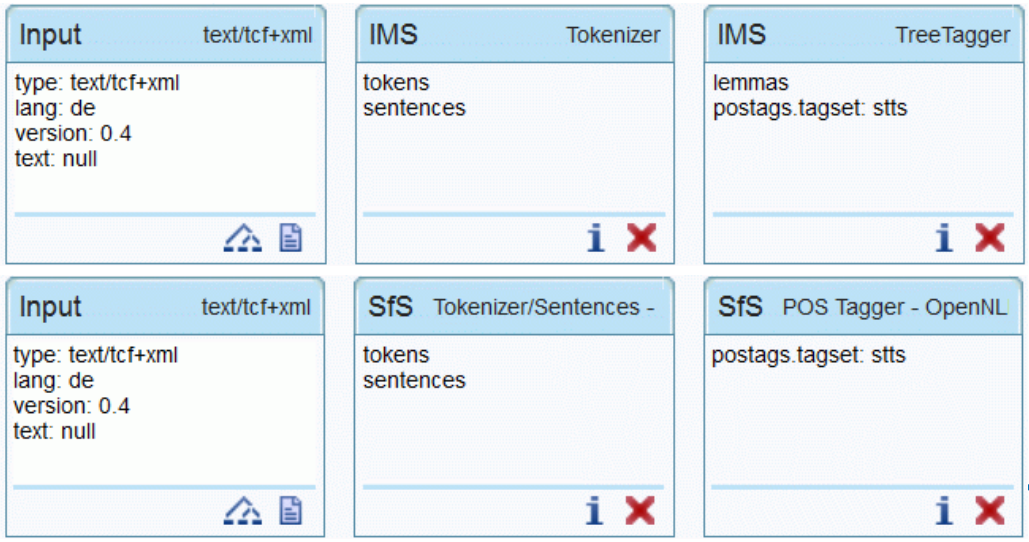
Phänomentyp	*Wikipedia-Diskussion*	*Chat*	*DWDS*
Verschriftete Umgangsprache I: Wortschreibung	20	20	(20)
IBK-typische oder nicht konventionalisierte Akronyme	20	20	
Verschriftete Umgangssprache II: Kontraktive Formen (VVFIN/VAFIN/VMFIN + PPER)	20	20	
IBK-spezifische Elemente I: Emoticons	20	20	
IBK-spezifische Elemente II: Aktionswörter	20	20	
Postings Gesamt:	**100**	**100**	
	200		

... analysiert mit zwei verschiedenen Toolchains in WebLicht, die beide das STTS-Tagset nutzen:

CLARIN-D | WEBLICHT WEB-BASED LINGUISTIC CHAINING TOOL

Toolchain 1: Kombinierter Tokenisierer und Satzgrenzenerkenner + TreeTagger des IMS

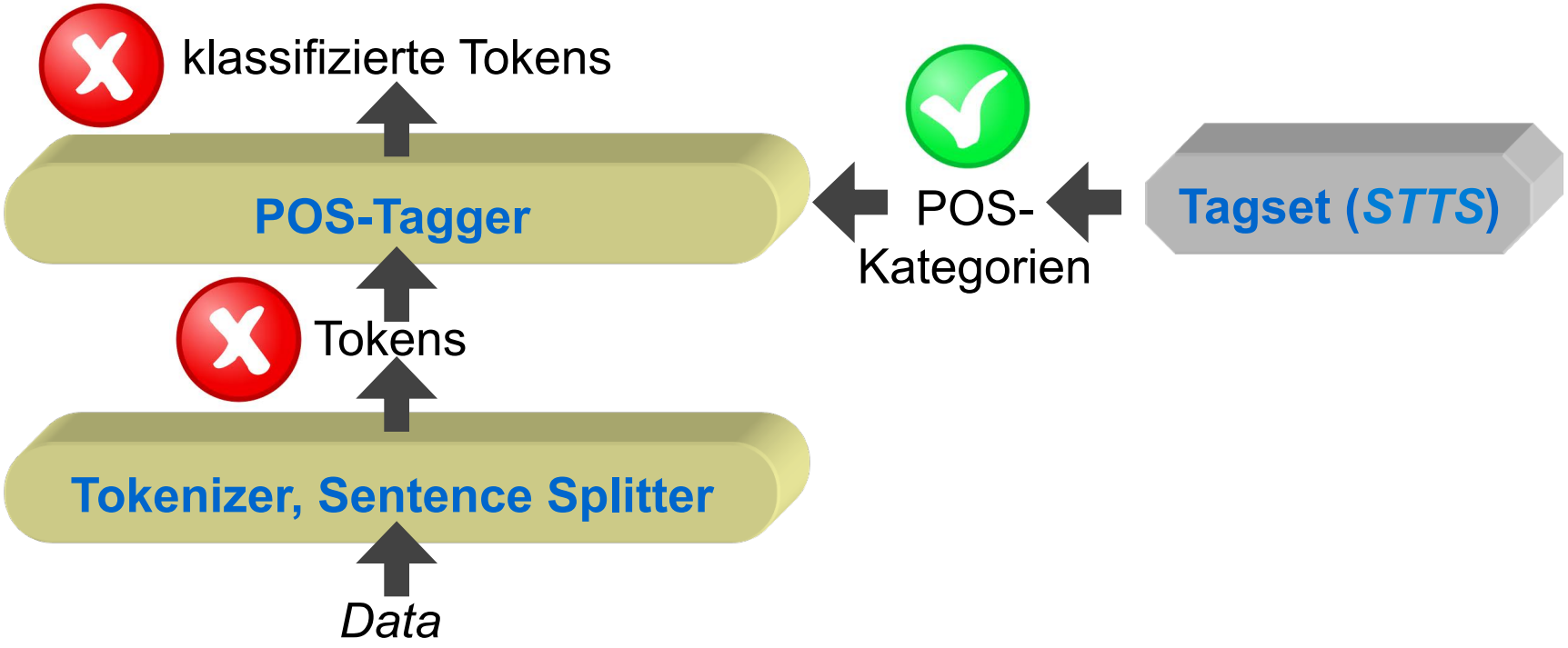
Toolchain 2: Kombinierter Tokenisierer und Satzgrenzenerkenner + Tagger aus dem OpenNLP-Projekt (SfS)



POS-Tagging für IBK-Daten: Probleme (und ihre Ursachen)

Problemtyp I: Tokenisierungs-Problem: Die Daten lassen sich in Tokens zerlegen, zu denen es Kategorien im Tagset gibt.
Der Tokenisierer liefert aber Tokens, die so segmentiert sind, dass sie sich nicht sinnvoll klassifizieren lassen.

⇨ Grund: Nicht-standardkonforme Verwendung von Spatien und Interpunktionszeichen (bei der Wort- und Satzschreibung).



POS-Tagging für IBK-Daten: Probleme (und ihre Ursachen)

Problemtyp I: Tokenisierungs-Problem: Die Daten lassen sich in Tokens zerlegen, zu denen es Kategorien im Tagset gibt.
Der Tokenisierer liefert aber Tokens, die so segmentiert sind, dass sie sich nicht sinnvoll klassifizieren lassen.

⇨ Grund: Nicht-standardkonforme Verwendung von Spatien und Interpunktionszeichen (bei der Wort- und Satzschreibung).

wieso stoeps?biste losgerannt einkaufen udn ahst vergessen dich anzuziehen vorher?*G*

Dortmunder Chat-Korpus, Dok. Nr. 2221007

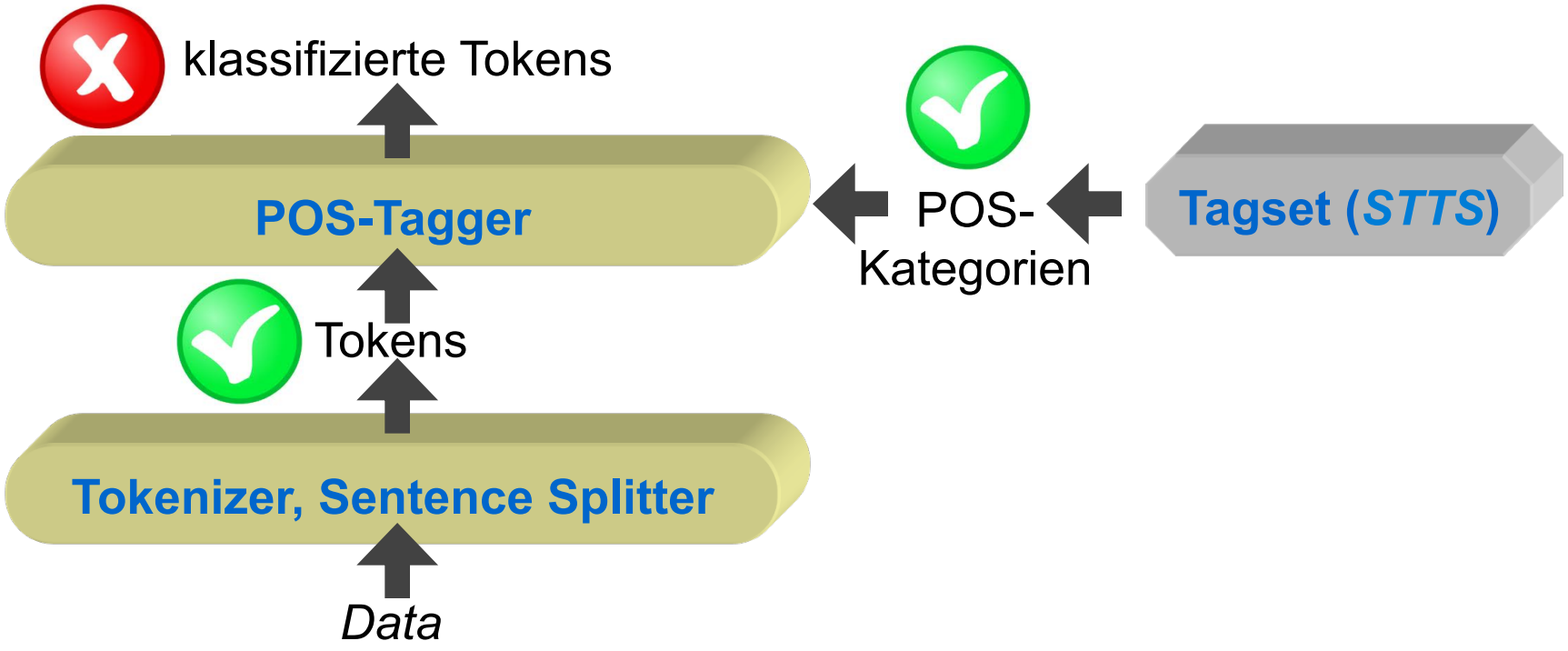
```
<token ID="t154">wieso</token>
<token ID="t155">stoeps?biste</token>
<token ID="t156">losgerannt</token>
<token ID="t157">einkaufen</token>
<token ID="t158">udn</token>
<token ID="t159">ahst</token>
<token ID="t160">vergessen</token>
<token ID="t161">dich</token>
<token ID="t162">anzuziehen</token>
<token ID="t163">vorher?*G*</token>
```

```
<token ID="t154">wieso</token>
<token ID="t155">stoeps</token>
<token ID="t156">?</token>
<token ID="t157">biste</token>
<token ID="t158">losgerannt</token>
<token ID="t159">einkaufen</token>
<token ID="t160">udn</token>
<token ID="t161">ahst</token>
<token ID="t162">vergessen</token>
<token ID="t163">dich</token>
<token ID="t164">anzuziehen</token>
<token ID="t165">vorher?*G*</token>
```

POS-Tagging für IBK-Daten: Probleme (und ihre Ursachen)

Problemtyp II: Klassifizierungs-Problem: Die Daten lassen sich in Tokens zerlegen, zu denen es Kategorien im Tagset gibt.
Der Tokenisierer segmentiert korrekt, der Tagger kann die Tokens aber nicht als Vertreter der vorhandenen Kategorien identifizieren.

⇨ tritt auf z.B. bei nicht-standardkonformen (an der umgangssprachlichen Lautung orientierten oder kreativen) Wortschreibungen sowie bei nicht konventionalisierten/okkasionellen Akronymen.



empirikom

Verschriftete Umgangssprache: Wortschreibung

DWDS	Wikipedia-Diskussionen	Chat
Ja, Meg Ryan habe tatsächlich ...	**Jut**, ich find die Variante mit ...	**jo**, mach das mal...
... **ja**, der Brownie koste zwei Dollar.	**Jo**, gute Vorbereitung ist ...	**japp** tom, stimmt. ...
... Prinzipiell **ja**, auch wenn ...	**Joh**, da hast Du sicher nicht ...	**jepp** zora, das bin ich ;)
... **ja**, in ihm offenbare sich der ...	**Jap**, geht klar!	**jau** das auto fährt ...
Goethe-Jahr? Aber **nein**: eine...	**Jupp**, aber Hinweise zu ...	**nope**,die 10000 gesamt sind ...
... dann sage ich: **nein**.	**Nee** dann müsste ich ja ...	@quaki, **nee**,bin ...
... **Nein**, nein: Der normale ...	Ach **nee**, jetze isses ...	**nöö** is er nich
... **Okay**, okay, sie ist ein ...	**Nö**, hat er nicht mehr ;-)	**nö**,dat ebste findeste ...
..., droht **jetzt** der Bankrott. ...	**Nööö** (Zitat Benutzer:Orientalist)...	**oki**...mach`s gut
Gut machst du **das**!, ruft ...	**okidoki**, sag Bescheid, wenn du ...	nö,**dat** ebste findeste ...
... gefroren ist, **das** ist schon ...	Ach nee, **jetze** isses plötzlich ...	**dat** ist donald duck
Als meine Augen wieder **gucken**...	Um Gottes Willen, geh **fott** mit ...	... einfach **guckst** was da ist ...
... Die Rose verblüht ihm **nicht**.	**Weia**, Augenkrebs hoch drei. ...	**tach** tomcat
..., in der Künstlerkolonie Worpswede **nicht** genug ...	**Tach** Wurm, geh mich doch ...	**nöö** is er nich
... mit dieser Spende **nichts** zu tun...	... geh mich doch **fott** mit ...	ich mag **net** wissen wie ...
... Bergtouren **nichts** anderes als ...	**Guck** Dir genau den kompletten ...	und sagt **nix**, der sack
Darum kann **ich** es ...	Hehe, **sowatt** kütt vüür ...	...kann man hier **nischt** mehr ...
..., wenn **ich** täglich einige...	Hehe, sowatt **kütt vüür** ...	... **isch** hab bestanden
... Mein richtiger **Vater** war ...	**Isch** ja gut, es hier noch ...	mach **isch** glatt :)
..., **aber** auch mit großem Aufwand...	... mit **Vadder** is hier Kim Il Sung...	ich auch **aba** bei mir ...

Verschriftete Umgangssprache: Wortschreibung

Korrekt klassifizierte Beleg-Instanzen...	*TreeTagger*	*POS Tagger OpenNLP*
... aus dem **DWDS-Korpus:**	**18** (20)	**15** (20)
... aus **Wikipedia-Diskussionsseiten:**	**1** (20)	**1** (20)
... aus **Chats:**	**2** (20)	**3** (20)

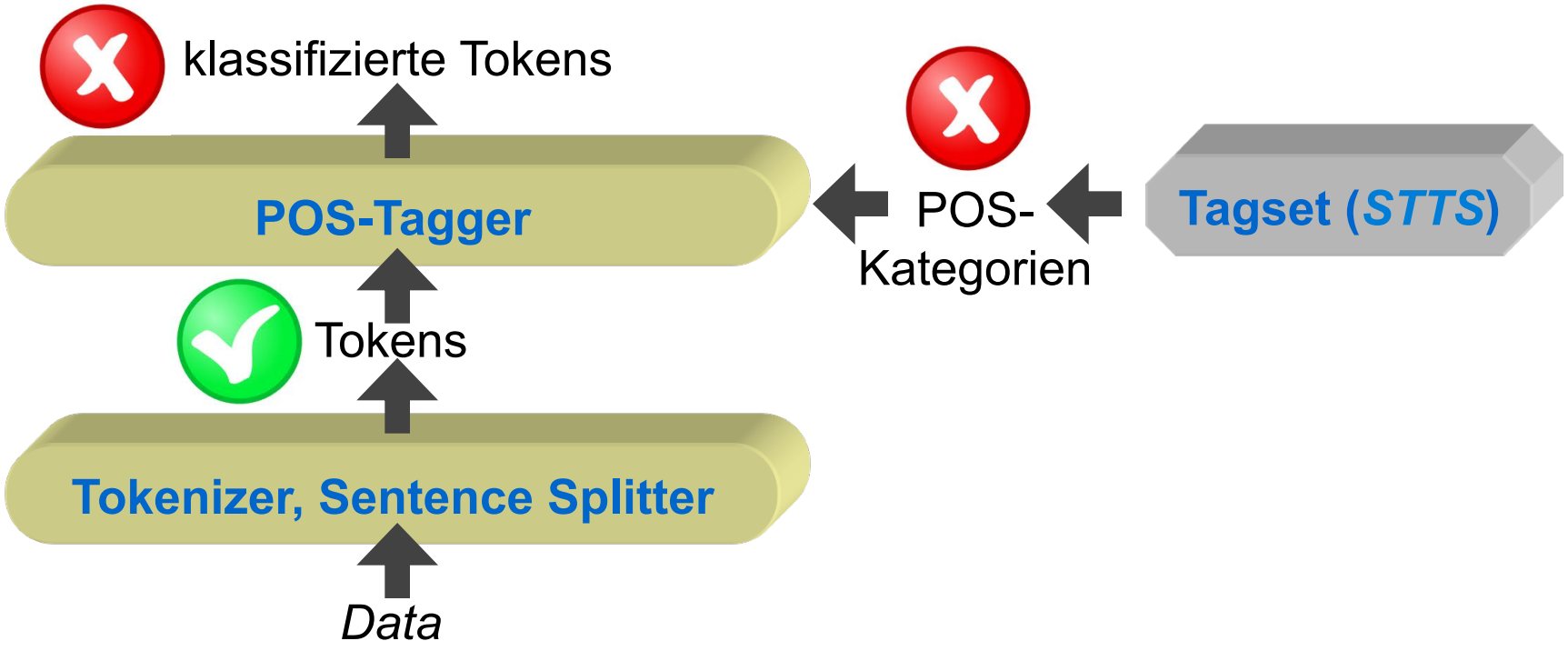
Vergleichbare Ergebnisse liefern die Tests zum Tagging IBK-typischer bzw. nicht-knoventionalisierter Akronyme:
IMHO, bspw., b.t.w., Btw., vllt, evt., mE, zB, Thx, jmd, LG, POV, ...

POS-Tagging für IBK-Daten: Probleme (und ihre Ursachen)

Problemtyp III: Kategorien-Problem: Der Tokenisierer segmentiert korrekt, der Tagger kann die Tokens aber nicht sinnvoll klassifizieren, da es im Tagset keine Kategorien dafür gibt.

⇨ tritt auf in folgenden Fällen:

(1) Tokens sind keine (oder keine prototypischen) *Wort*-Tokens;
(2) Tokens gehören zu Kategorien, die erst noch an existierende POS-Einteilungen anzubinden sind.



Verschriftete Umgangssprache: Kontraktive Formen

Beispiel: **Kontraktive Formen des Typs VVFIN / VAFIN / VMFIN + PPER (+ PPER)**

haste, biste, findeste, könnteste, magste, meinste, denkste, machste, machstes, isses, hats, kanns, kenns, gehts, habs, sags, schreibs, machs, machts, wärs, wirds

Beleg-Instanzen klassifiziert als...	Wikipedia-Diskussionen:		Chat:	
	TreeTagger	*OpenNLP*	*TreeTagger*	*OpenNLP*
VVFIN / VAFIN:	**8** (20)	**7** (20)	**7** (20)	**10** (20)
NN:	**8** (20)	**1** (20)	**6** (20)	**0** (20)
ADJA / ADJD:	**4** (20)	**5** (20)	**7** (20)	**4** (20)
ADV:	**0** (20)	**0** (20)	**0** (20)	**3** (20)
andere:	**0** (20)	**7** (20)	**3** (20)	**3** (20)

IBK-typische Stilelemente

:) (: :-)
:-)) :-)))
:o) ;-)
;-))))
:(:-(:O)
:P :-P 8)
=o) ^^ -.-
o_O O-O

freu *lach*
lächel *grins*
fiesgrins
wink Gähn
Seufz *werb*
wunder
stotter
rotwerd
einrück lol
LOL *lol* *rofl*
Grummel
kopfschüttel
duck *g*
ggg
lernenmuss

Emoticons:

(a) Tokenisierungs-Problem: 27 von 40 Emoticon-Tokens inkorrekt segmentiert (IMS-Tokenisierer + TreeTagger)

```
<token ID="t99">:-</token>
<token ID="t100">)</token>
<token ID="t101">)</token>
```

(b) Tagging-Problem (auch bei manuell normalisierter Tokenisierung): Keine Kategorie im Tagset (STTS) passt:

POS-Tag	Anzahl
NN	20
ADJD	12
ADJA	6
NE	1
VVFIN	1
Gesamt	**40**

Inflektive:

(a) Tokenisierungs-Problem: Asterisken werden meist nicht vom sprachlichen Token abgetrennt.

(b) Tagging-Problem (auch bei manuell normalisierterTokenisierung): Keine Kategorie im Tagset passt:

POS-Tag	Anzahl
NN	21
VVIMP	9
NE	3
ADJA	5
ADJD	1
PTKVZ	1
Gesamt	**40**

POS-Tagging für IBK-Daten: Herausforderungen

Um das POS-Tagging für IBK-Daten zu verbessern, müssen Taggingverfahren auf verschiedenen Ebenen angepasst werden:

- auf der Ebene der Tokenisierung und Satzgrenzenerkennung;
- auf der Ebene des Taggings (Zuordnung von Tokens zu Kategorien im Tagset);
- auf der Ebene der Tagsets selbst (Modifikation und Erweiterung von Kategorien für die Erfassung semiotischer und sprachlicher Innovationen in IBK).

IBK Tagging ≠ Standard Tagging

Spending the day withhh mommma !

Spending	the	day	withhh	mommma	!
V	D	N	**P**	N	,

Twitter Tokenizer +
Twitter Tagger

Spending	the	day	withhh	mommma	!
VVG	DT	NN	**NN**	NNS	SENT

Standard Tokenizer +
TreeTagger

IBK Tagging ≠ Standard Tagging

Spending the day withhh mommma !

Vereinfachtes Tagset

Spending	the	day	withhh	mommma	!
V	D	N	**P**	N	,

Twitter Tokenizer +
Twitter Tagger

Spending	the	day	withhh	mommma	!
VVG	DT	NN	**NN**	NNS	SENT

Standard Tokenizer +
TreeTagger

IBK Tagging ≠ Standard Tagging

Spending the day withhh mommma !

	Spending	the	day	withhh	mommma	!
Twitter Tokenizer + Twitter Tagger	V	D	N	**P**	N	,
	Spending	the	day	withhh	mommma	!
Standard Tokenizer + TreeTagger	VVG	DT	NN	**NN**	NNS	SENT

Angepasstes Modell → Besseres Tagging

IBK Tagging ≠ Standard Tagging

Spending the day withhh mommma !

Generalisierte POS-Tags

Spending	the	day	withhh	mommma	!
V	D	N	P	N	,
V	ART	NN	PP	NN	PUNC

Twitter Tokenizer + Twitter Tagger

Spending	the	day	withhh	mommma	!
VVG	DT	NN	NN	NNS	SENT
V	ART	NN	NN	NN	PUNC

Standard Tokenizer + TreeTagger

Einfluss der Tokenisierung

Different smiley styles :) :-) (^_^) ^o #smiley

Different	smiley	styles		:)			:-)	(^			^)		^o		#smiley	
A	A	N		E			E	E			EMO		EMO		HASH	
ADJ	ADJ	NN		EMO			EMO	EMO			E		E		#	
Different	smiley	styles	:	)	:	-	)	(	^	_	^	)	^	o	#	smiley
JJ	NN	NNS	:	)	:	:	)	(	SYM	SYM	SYM	)	SYM	NN	#	NN
ADJ	NN	NN	O	O	O	O	O	O	O	O	O	O	O	NN	O	NN

Einfluss der Tokenisierung

Different smiley styles :) :-) (^_^) ^o #smiley

Unterschiedliche Tokenisierung

Different	smiley	styles	:)	:-)	(^	^)	^o	#smiley
A	A	N	E	E	E	EMO	EMO	HASH
ADJ	ADJ	NN	EMO	EMO	EMO	E	E	#

Different	smiley	styles	:	)	:	-	)	(	^	_	^	)	^	o	#	smiley
JJ	NN	NNS	:	)	:	:	)	(	SYM	SYM	SYM	)	SYM	NN	#	NN
ADJ	NN	NN	O	O	O	O	O	O	O	O	O	O	O	NN	O	NN

RT	@torsten_zesch	I	cannot	belive	(	sic	)	how	well	this	tagger	works	:)	http://code.google.com/jwpl/
~	@	O	V	N	,	R	,	R	R	D	N	V	E	U
DM	AT	PR	V	NN	PUNC	ADV	PUNC	ADV	ADV	ART	NN	V	EMO	URL

RT	@	torsten_zesch	I	cannot	belive	(	sic	)	how	well	this	tagger	works	:	)	http	:	/	/	code.google.com	/	jwpl	/
NP	SYM	NN	PP	MD	VV	(	JJ	)	WRB	RB	DT	NN	VVZ	:	)	NN	:	SYM	SYM	NN	SYM	NN	SYM
NP	O	NN	PR	V	V	O	ADJ	O	ADV	ADV	ART	NN	V	O	O	NN	O	O	O	NN	O	NN	O

Spezialisierte Tagsets

Different smiley styles :) :-) (^_^) ^o #smiley

Different	smiley	styles	:)	:-)	(^	^)	^o	#smiley
A	A	N	E	E	E	EMO	EMO	HASH
ADJ	ADJ	NN	EMO	EMO	EMO	E	E	#

Different	smiley	styles	:	)	:	-	)	(	^	_	^	)	^	o	#	smiley
JJ	NN	NNS	:	)	:	:	)	(	SYM	SYM	SYM	)	SYM	NN	#	NN
ADJ	NN	NN	O	O	O	O	O	O	O	O	O	O	O	NN	O	NN

Spezialisiertes Twitter-Tagset

Einblicke in die empirikom-Werkstatt

Zoom-in in drei Bereiche gegenwärtiger Forschung:

- POS-Tagging für internetbasierte Schriftlichkeit: Problemaufriss auf Basis von Experimenten mit Wikpedia-, Chat- und Twitter-Daten
- Verarbeitung von IBK-Korpora / Webkorpora: Herausforderungen und Lösungsansätze
- Experimente zur linguistischen Evaluation von Webkorpora

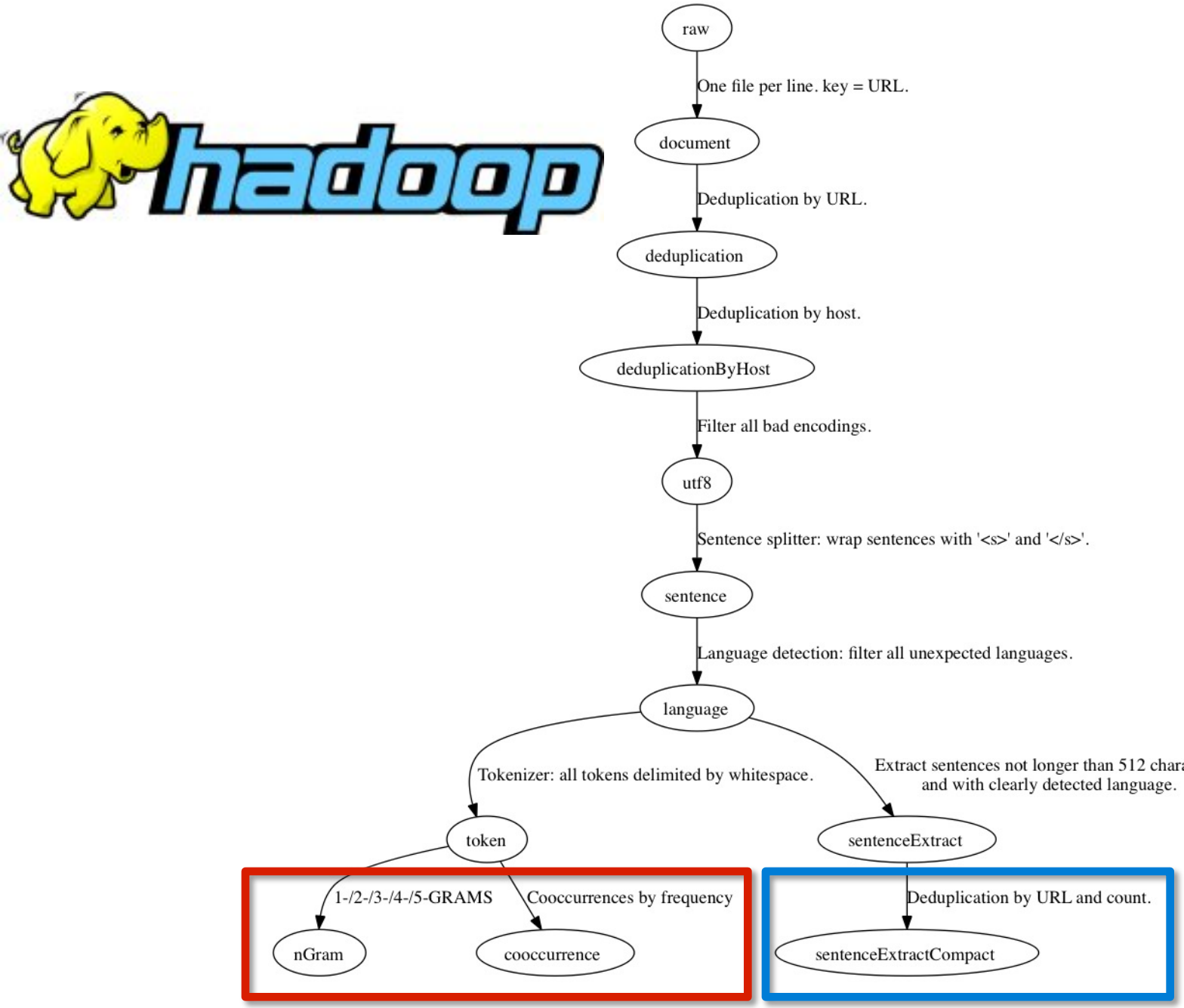
Ausgangslage

- Gewinnung

- Rechtliche Aspekte / Archivierung

→Verarbeitung großer Korpora?

(Zusammen mit Chris Biemann)


hadoop
raw
One file per line. key = URL.
document
Deduplication by URL.
deduplication
Deduplication by host.
deduplicationByHost
Filter all bad encodings.
utf8
Sentence splitter: wrap sentences with '<s>' and '</s>'.
sentence
Language detection: filter all unexpected languages.
language
Tokenizer: all tokens delimited by whitespace.
Extract sentences not longer than 512 characters and with clearly detected language.
token
sentenceExtract
1-/2-/3-/4-/5-GRAMS
Cooccurrences by frequency
Deduplication by URL and count.
nGram
cooccurrence
sentenceExtractCompact

Statistiken

- 8-node Hadoop-Cluster mit 70 map slots und 35 reduce slots

Job	**4GB** minutes	GB	**40GB** minutes	GB
Raw	-	4	-	40
Document	2	4,24	15	42,36
Deduplication	23	3,81	105	20,81
DeduplicationByHost	9	1,56	53	12,21
UTF8	1	1,54	3	10,63
Sentence	28	1,61	130	11,12
Language	5	1,86	23	12,31
Token	5	1,91	24	12,63
NGram (1)	5	0,07	19	0,21
NGram (2)	5	0,68	21	2,27
NGram (3)	5	2,08	21	7,87
NGram (4)	5	3,71	22	15,27
NGram (5)	5	5,19	24	22,43
Cooccurrence (5)	8	6,54	30	22,91
SentenceExtract	4	1,62	19	10,83
SentenceExtractCompact	1	1,43	3	6,98
Total	111	61,91	512	210,83

Tiefe Automatische Annotation

- Beliebige UIMA/DKPro Komponenten zusätzlich ausführen

http://code.google.com/p/dkpro-core-asl/

- „tiefe“ automatische Annotation der Korpora
  - Diverse Tagger-Implementierungen
  - 3 Dependenzparser mit deutschen Modellen (+1 inoffiziell)
  - Semantic Role Labeling
  - Koreferenzauflösung
  - Lexikalische Ketten
  - Lesarten- / Eigennamendisambiguierung
  - …

Offen Fragen

- Sind Webkorpora ein Ersatz für traditionelle Referenzkorpora?

- „bigger is better“?

Einblicke in die empirikom-Werkstatt

Zoom-in in drei Bereiche gegenwärtiger Forschung:

- POS-Tagging für internetbasierte Schriftlichkeit:
  Problemaufriss auf Basis von Experimenten mit Wikpedia-, Chat- und Twitter-Daten
- Verarbeitung von IBK-Korpora / Webkorpora:
  Herausforderungen und Lösungsansätze
- Experimente zur linguistischen Evaluation von Webkorpora

Evaluation von Webkorpora: POS-Tagging

Webkorpora als Ersatz für traditionelle Referenzkorpora

- relativ nah am schriftsprachlichen Standard ≠ IBK-Korpora
  - herkömmliche Analysekategorien und -werkzeuge anwendbar
- wesentlich größer (→ höhere Abdeckung, besser für Statistik)
- linguistische Annotation möglich ≠ Google-Onlinesuche

Evaluation von POS-Tagging für deutsche Web-Korpora

- POS-Tagging Voraussetzung für computerlinguistische Auswertung
- Masterarbeit von Eugenie Giesbrecht (empirikom):
  manuelle Annotation von Stichproben aus deWaC-Webkorpus
  - STTS-Tagset und Tokenisierungskriterien
- Vergleich mit TreeTagger, TnT-Tagger, Stanford Tagger, …
- Richtwert: Genauigkeit auf Zeitungstexten ca. 97%–98%

Evaluation von Webkorpora: POS-Tagging

	Genre	TT-SPF[a)]	TT[b)]	TnT	Stanford	SVM	UIMA
1.	*TV episode guide*	***93.89***	*90.87*	*92.79*	*92.83*	*92.78*	*89.91*
2.	news report (medicine)	**96.88**	97.12	95.92	96.16	95.68	94.26
3.	political speech	**97.52**	96.56	96.42	96.15	93.81	95.61
4.	job market news	**97.46**	93.65	96.19	96.95	95.18	95.44
5.	story (Paul of Thebes)	**95.42**	94.84	95.08	95.37	95.08	93.87
6.	exposition programme	**94.23**	92.13	92.83	92.66	93.01	90.75
7.	*online forum*	***88.01***	*79.97*	*85.56*	*84.47*	*84.51*	*84.47*
8.	report on infections	**98.25**	96.89	97.28	**98.25**	97.08	95.54
9.	*conference information*	*90.98*	*89.18*	*92.01*	*90.98*	***93.30***	*92.55*
10.	IT news (CeBIT)	93.69	92.73	92.93	94.07	94.07	**95.42**
11.	info (support programme)	97.10	98.51	98.01	**99.50**	97.01	98.02
12.	*news report (archbishop)*	*91.97*	*87.15*	*91.97*	*91.97*	***93.97***	*90.80*
13.	synopsis of cold war	96.67	94.86	96.49	95.68	95.40	**97.30**
		94.77 ±3.04	92.65 ±5.04	94.11 ±3.31	94.23 ±3.85	93.91 ±3.15	93.38 ±3.67

(Giesbrecht & Evert 2009)

Linguistische Evaluation von Webkorpora

Fragestellung

- Webkorpora geeignet als Ersatz für traditionelle Referenzkorpora in Korpuslinguistik und Sprachtechnologie?
- Größe → bessere Ergebnisse bei quantitativer Auswertung?

Ansätze zur Evaluation

- direkter Vergleich von Häufigkeitsdaten mit Referenzkorpus (z.B. Baroni *et al.* 2009, Keller & Lapata 2003)
- inhaltliche Beurteilung (Ähnlichkeit zu Referenzkorpus)
- Vergleich mit Plausibilitätsurteilen (Keller & Lapata 2003)
- Einsatz in computerlinguistischer Anwendung (z.B. Turney 2001)
- Identifikation von Kollokationen und Mehrwortausdrücken (MWE)
- Basis für distributionelle Semantik (*wordspace models*, DSM)

Hier: Kollokationsextraktion (am interessantesten)

Verwendete Korpora

- **British National Corpus** (BNC, 100 M)
  - Referenzkorpus als Basis für Vergleiche
- **Wikipedia** (Wackypedia: ca. 1 G / WP500: 200 M)
  - „fast traditionelles“ Korpus (Enzyklopädie)
- **ukWaC** Web Corpus (ca. 2 G)
  - weit verbreitetes Webkorpus, heuristisch bereinigt
  - POS-Tagging, Lemmatisierung, Dependenz-Parsing
- Google **Web 1T 5-Grams** (ca. 1 T = 1000 G!)
  - minimal bereinigt, keine linguistische Annotierung
  - nur als N-Gramm-Datenbank (max. 5-Gramme mit $f \geq 40$)
  - „Quasi-Kollokationen“ = Kookkurenzinformation aus N-Grammen
- **Leipzig Corpora Collection** (LCC, ca. 1.5 G)
  - automatisch bereinigt, (noch) keine linguistische Annotierung
  - vollständige N-Gramme und Kookkurrenzdaten (Darmstadt)
  - Vergleich Kollokationen / Quasi-Kollokationen

Linguistische Evaluation: Kollokationsextraktion

Kollokationsextraktion mit Assoziationsmaßen

- Liste von Kandidaten (Wortpaaren) aus Korpus extrahiert
- diverse Filter (Wortarten, syntaktische Muster, Häufigkeit)
- Kandidaten anhand statistischer Assoziationsmaße sortiert
  → am höchsten bewertete Kandidaten werden manuell gesichtet

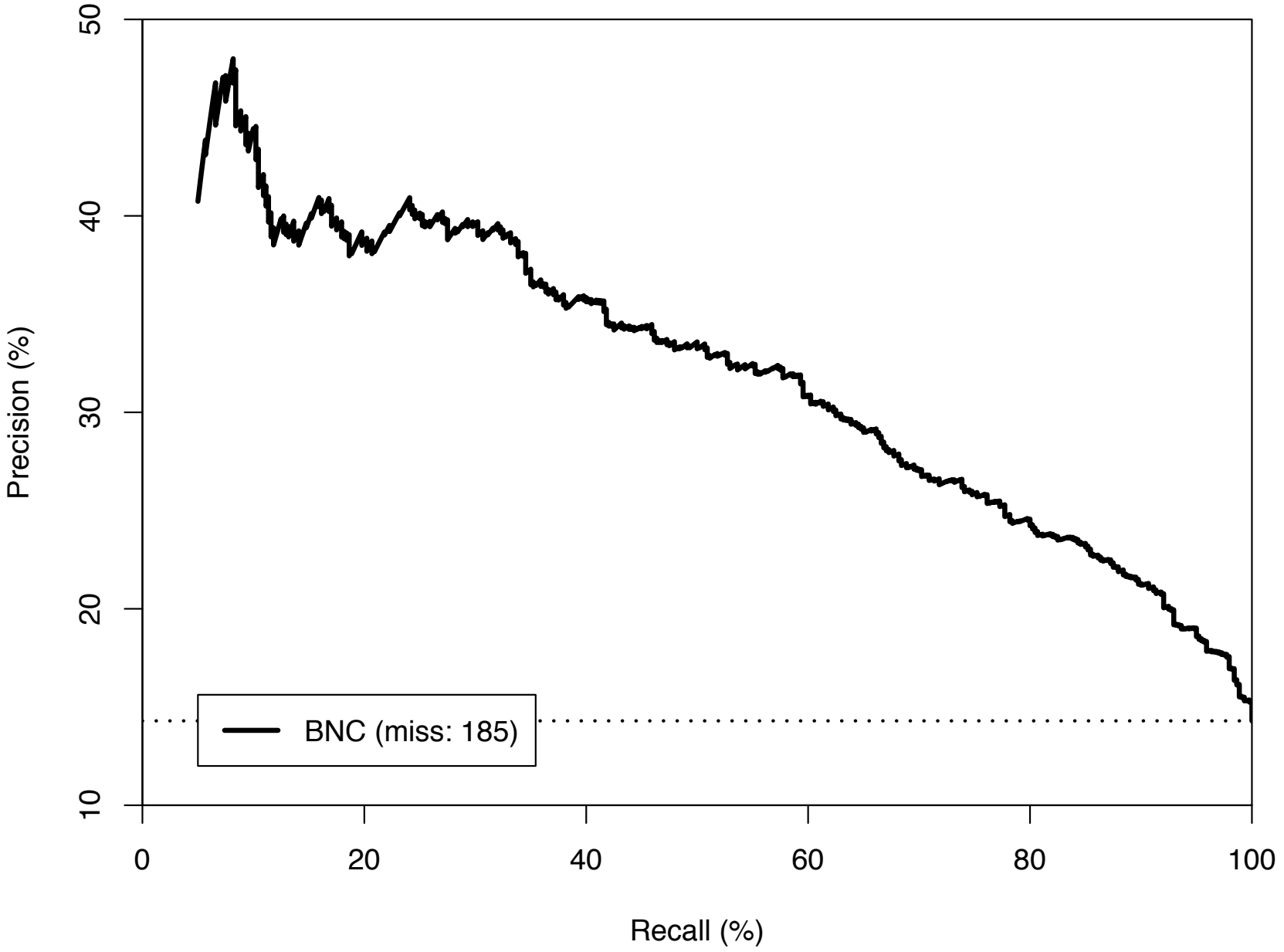
Evaluation der Assoziationsmaße

- Kandidaten manuell annotiert: True vs. False Positives (TP / FP)
- für *n* am höchsten bewertete Kandidaten (*n-best* list) werden *Precision* und *Recall* berechnet
- grafische Darstellung: *Precision* gegen *n-best* bzw. *Recall*

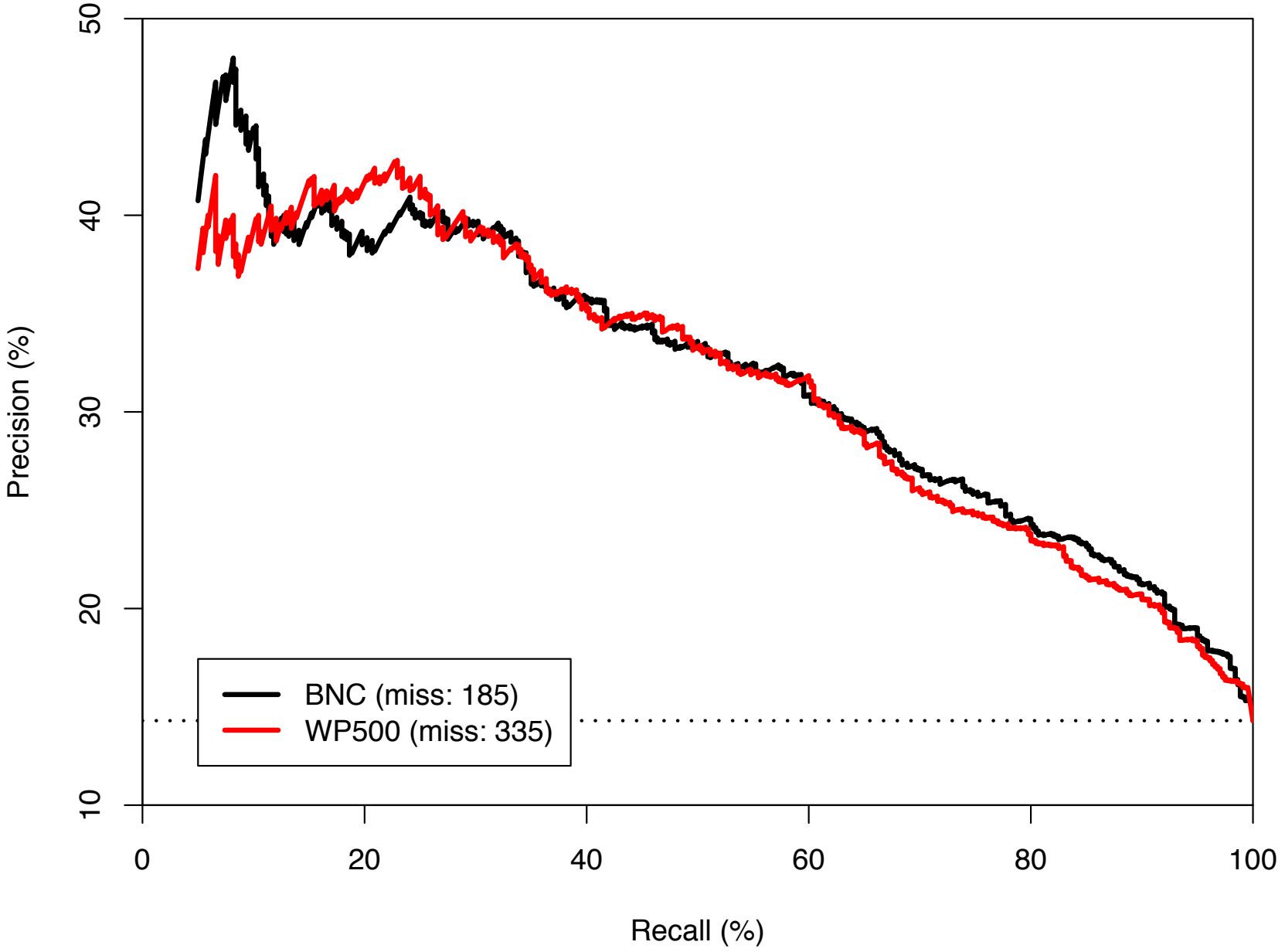
Referenzdaten (*Gold Standard*) für Englisch

- Partikelverben (VPC, Baldwin 2008) – 440 / 3078 Kandidaten
- Funktionsverbgefüge (LVC, Tu & Roth 2011) – 349 / 891 Kandidaten
- lexikalische Kollokationen (BBI, Benson *et al.* 1986) – 36328 TPs

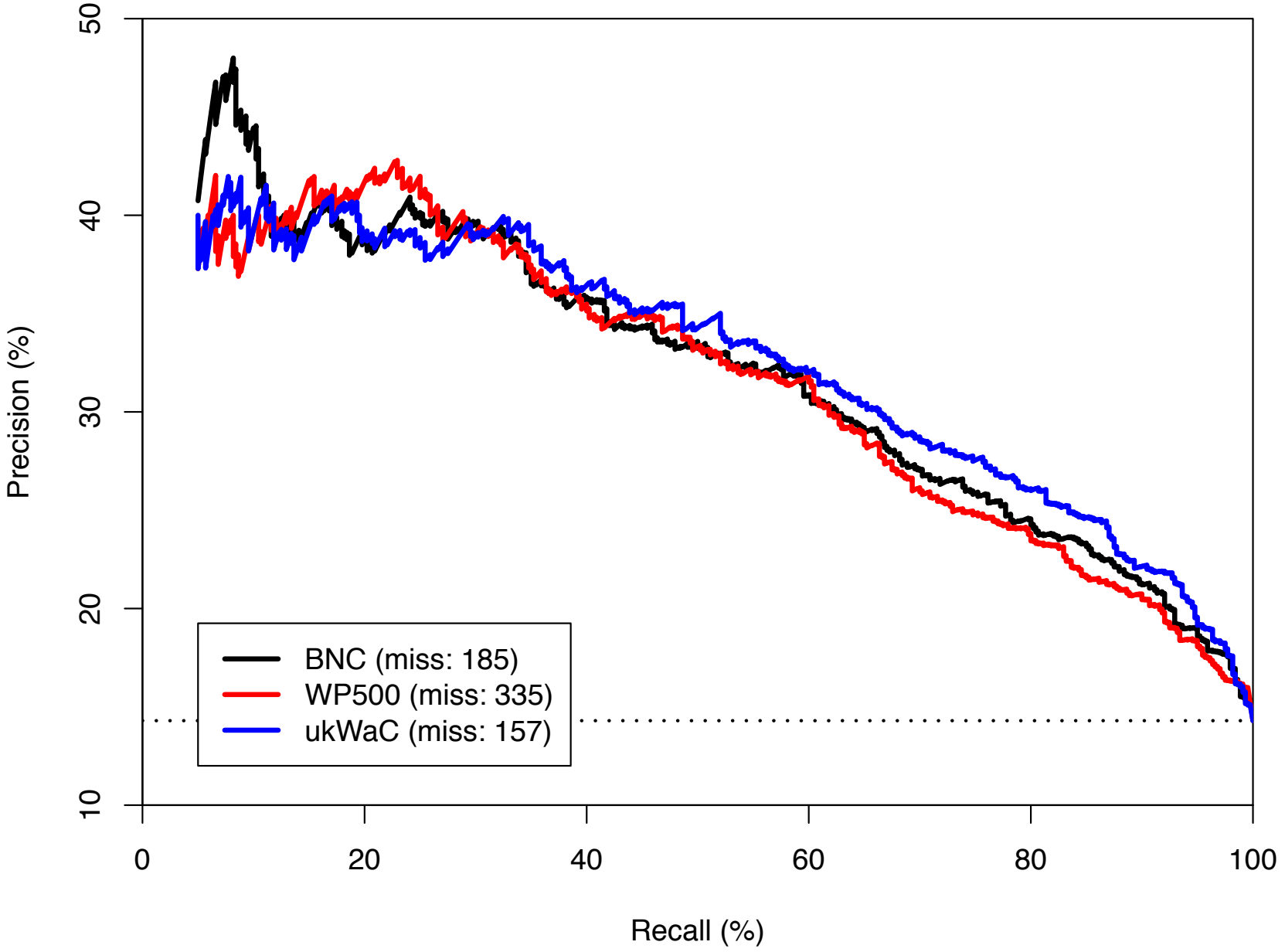
Linguistische Evaluation: VPC



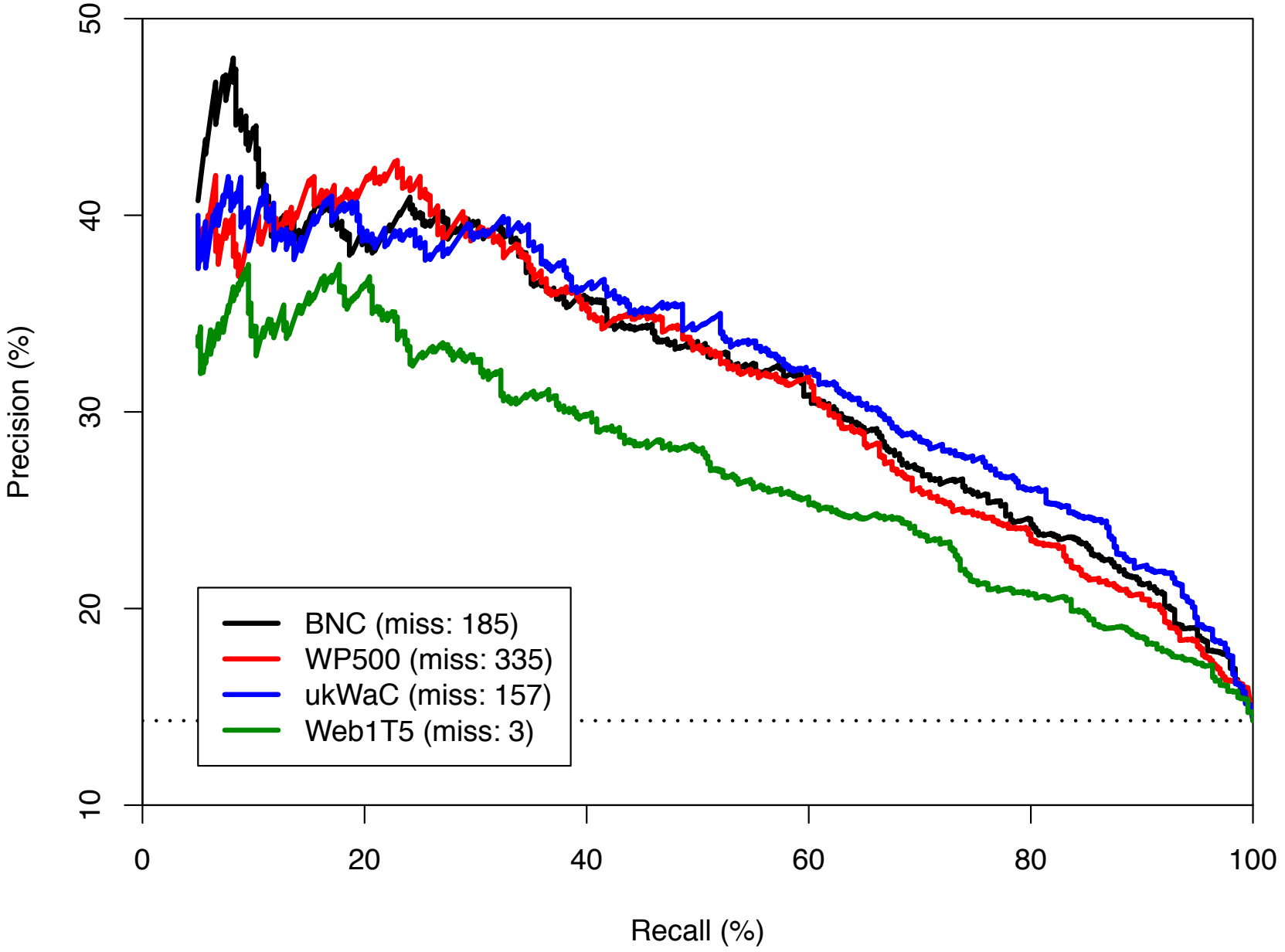
Linguistische Evaluation: VPC



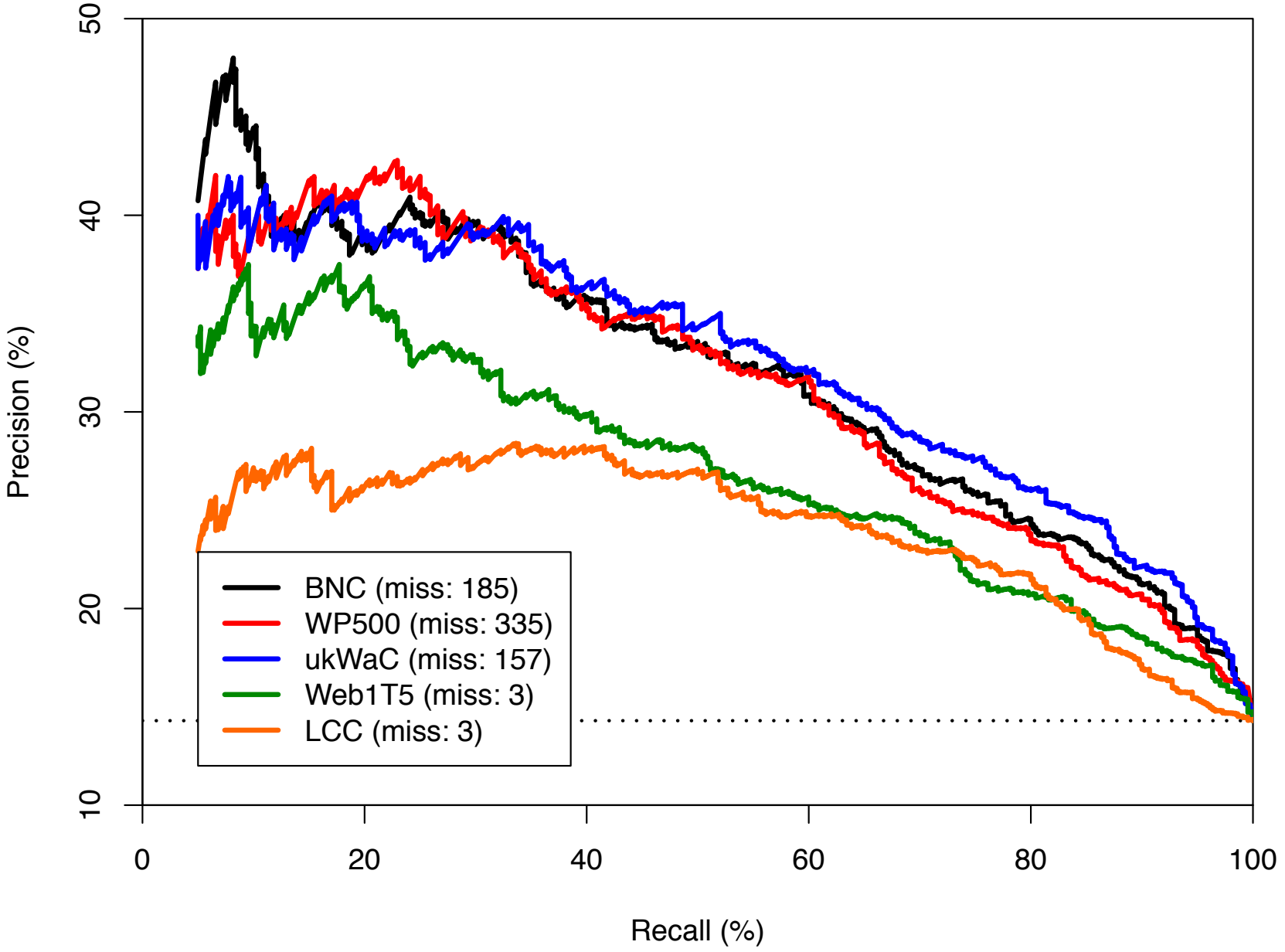
Linguistische Evaluation: VPC



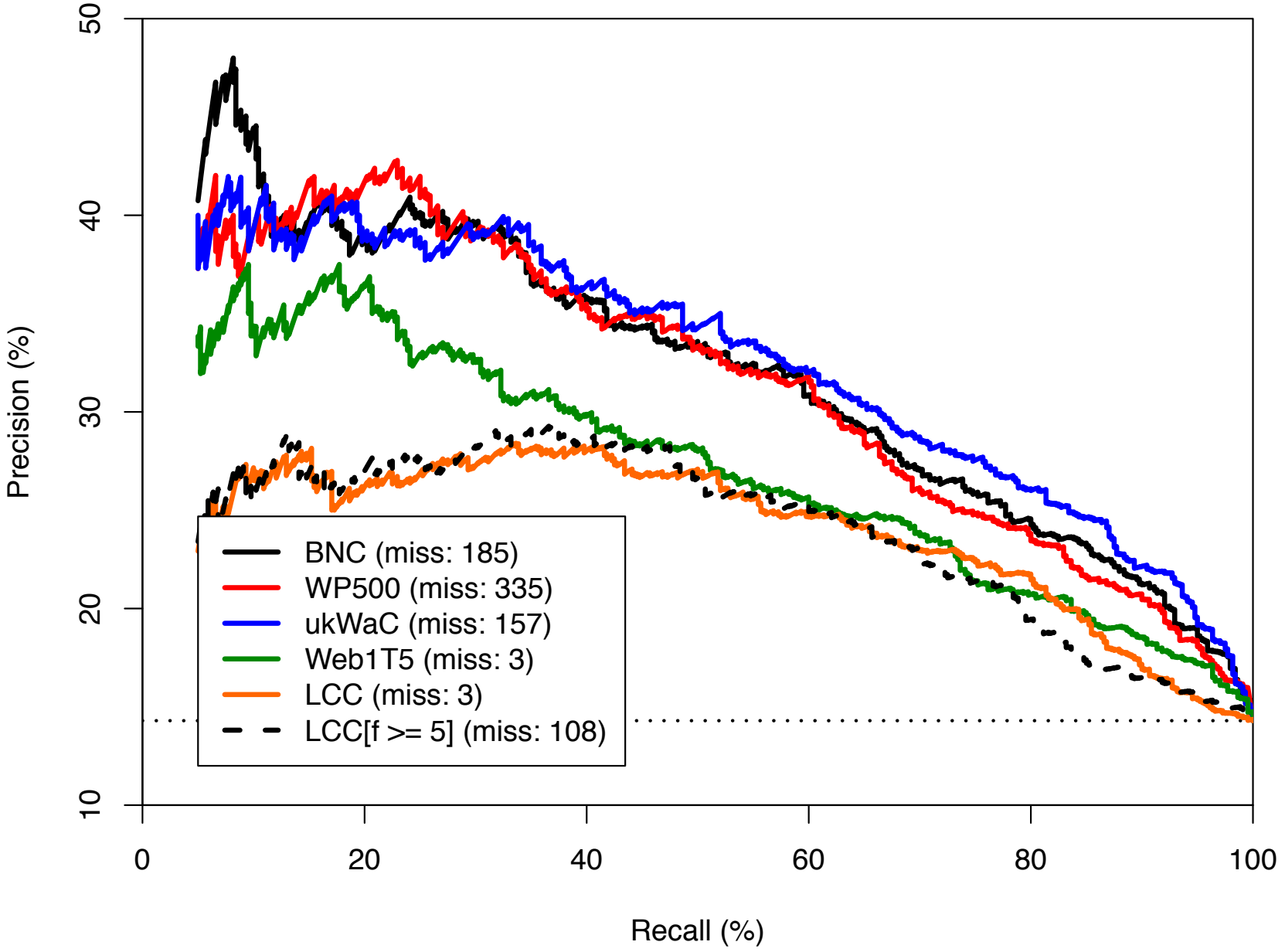
Linguistische Evaluation: VPC



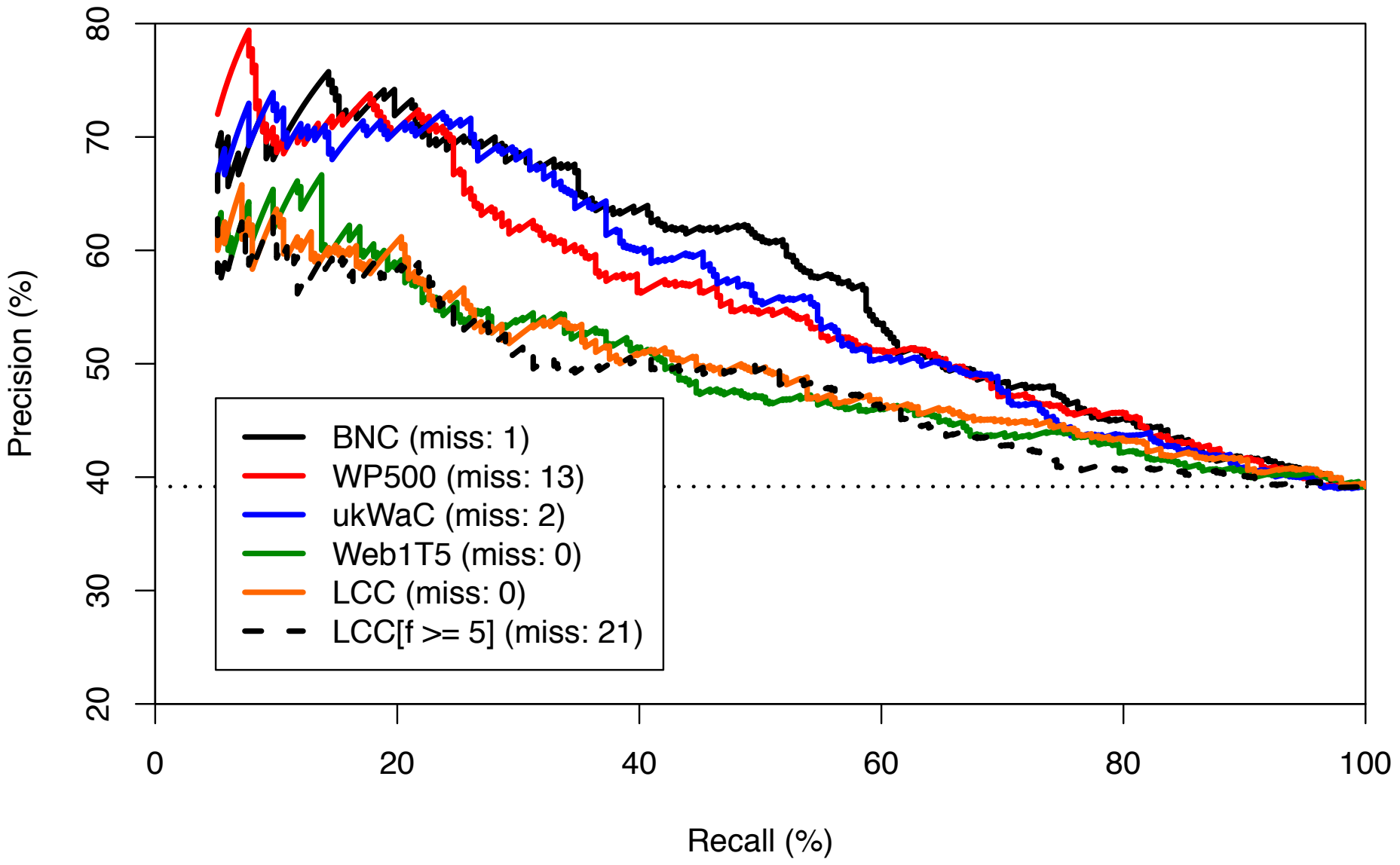
Linguistische Evaluation: VPC



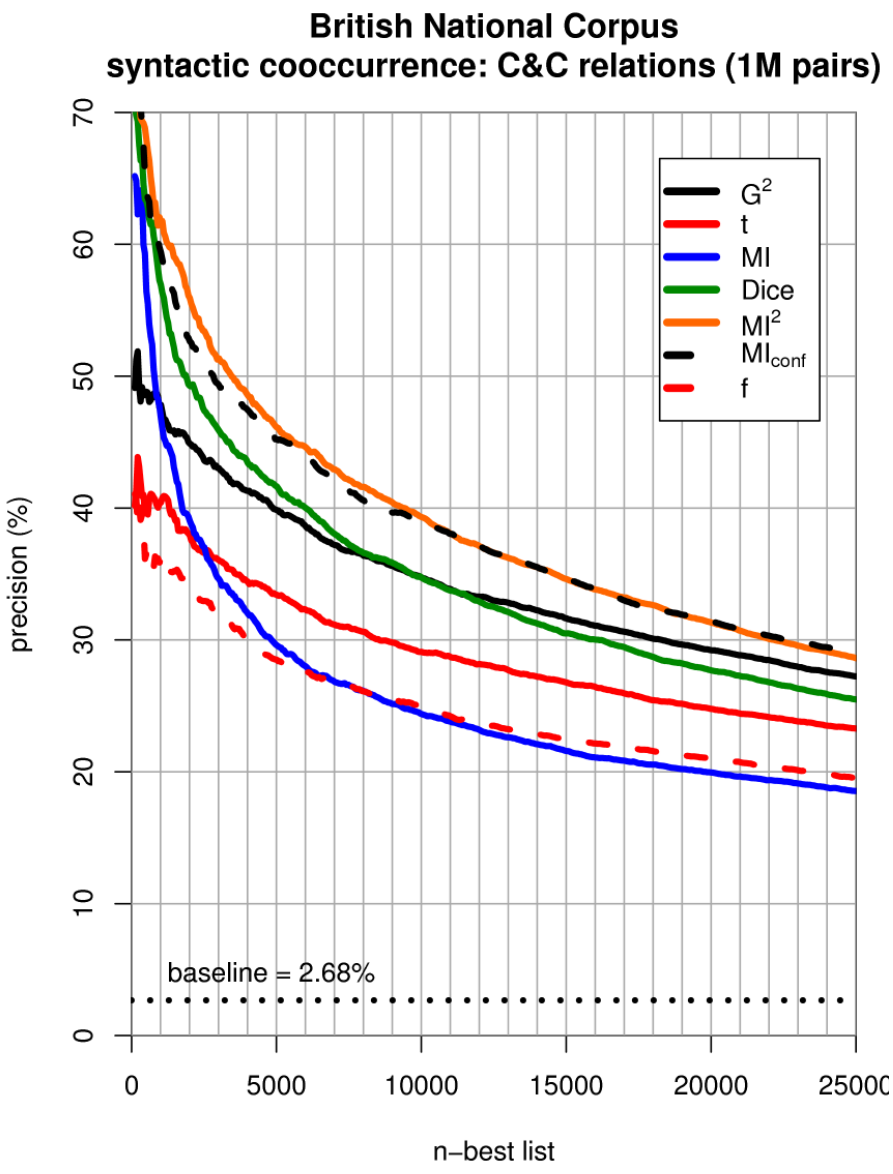
Linguistische Evaluation: VPC

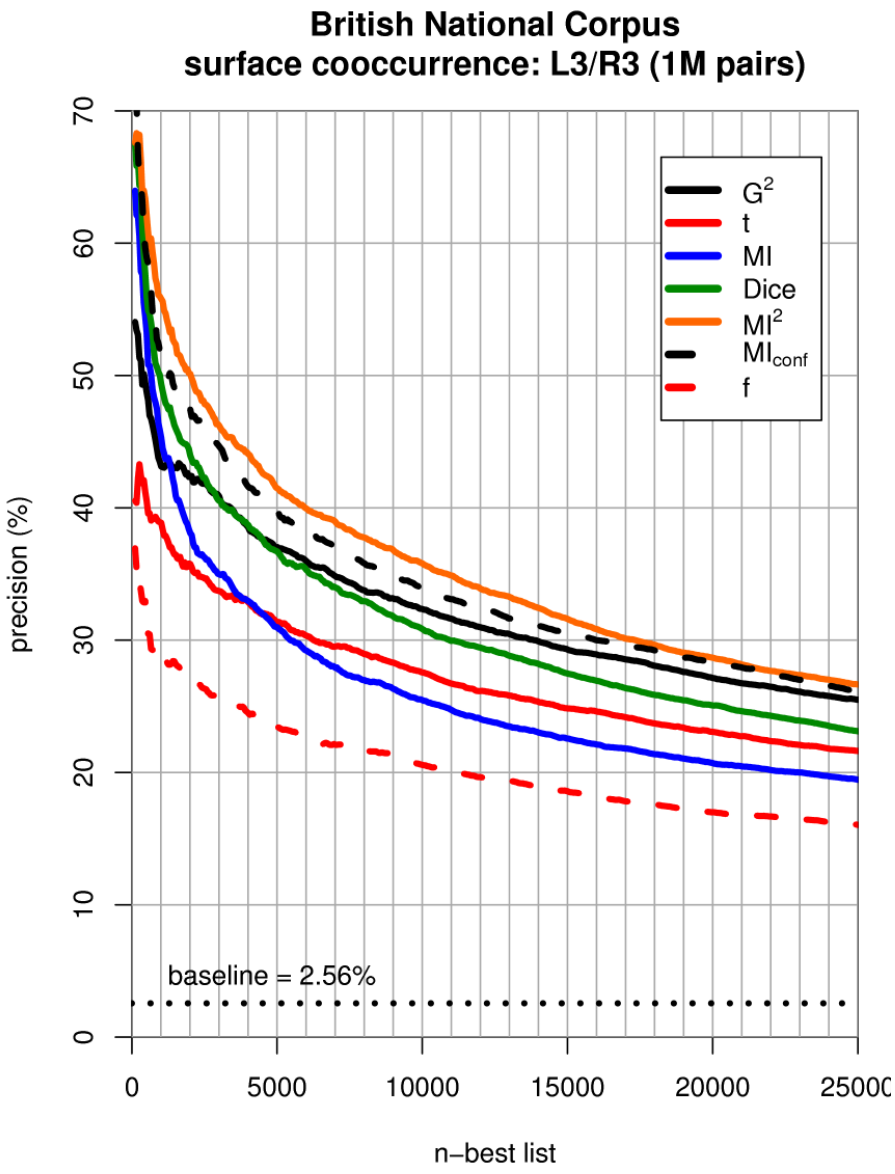


Linguistische Evaluation: LVC

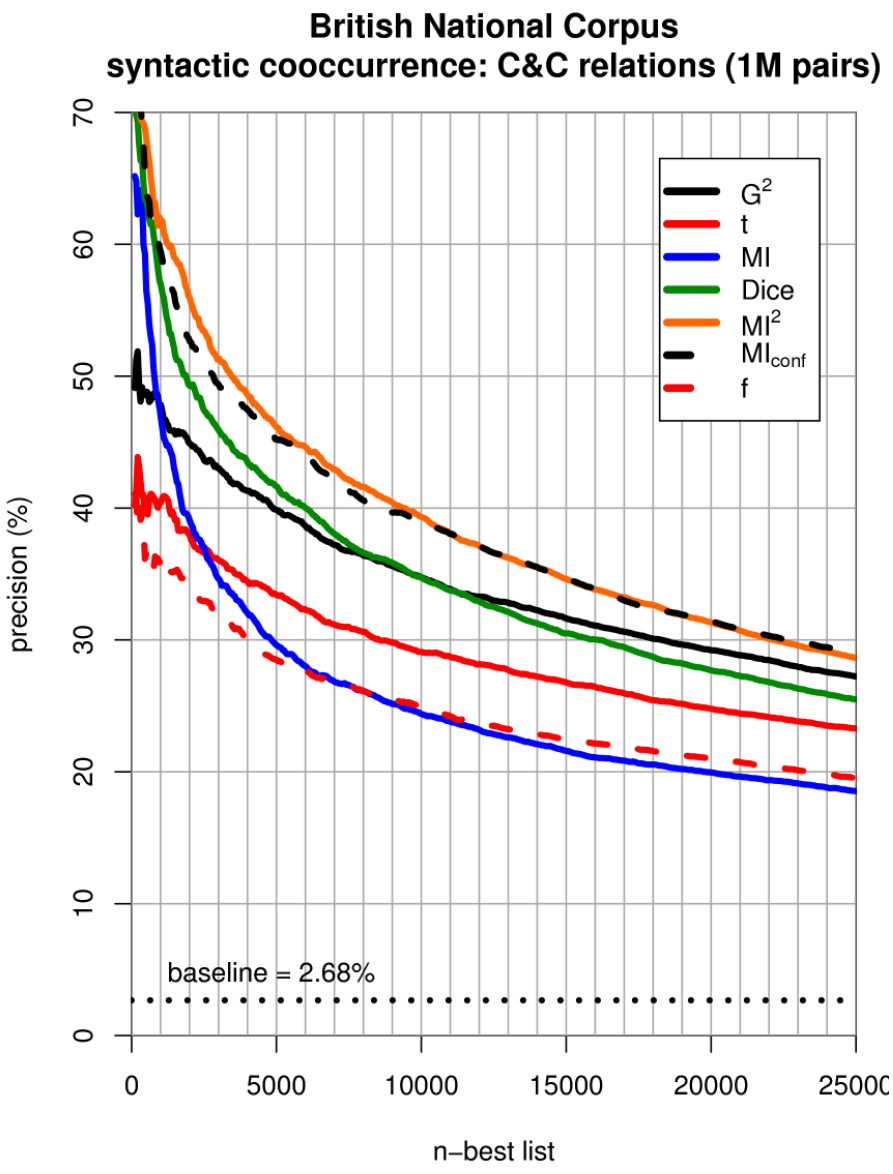


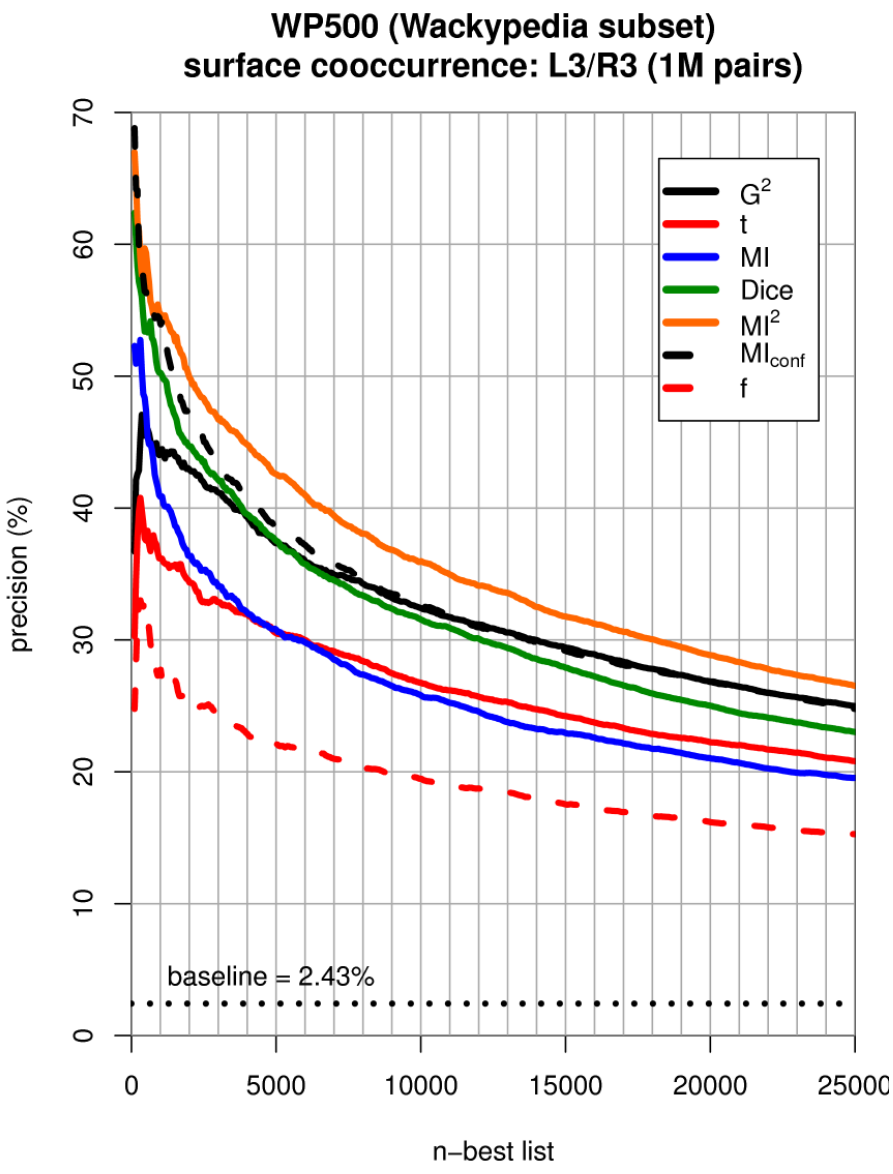
Linguistische Evaluation: BBI-Kollokationen



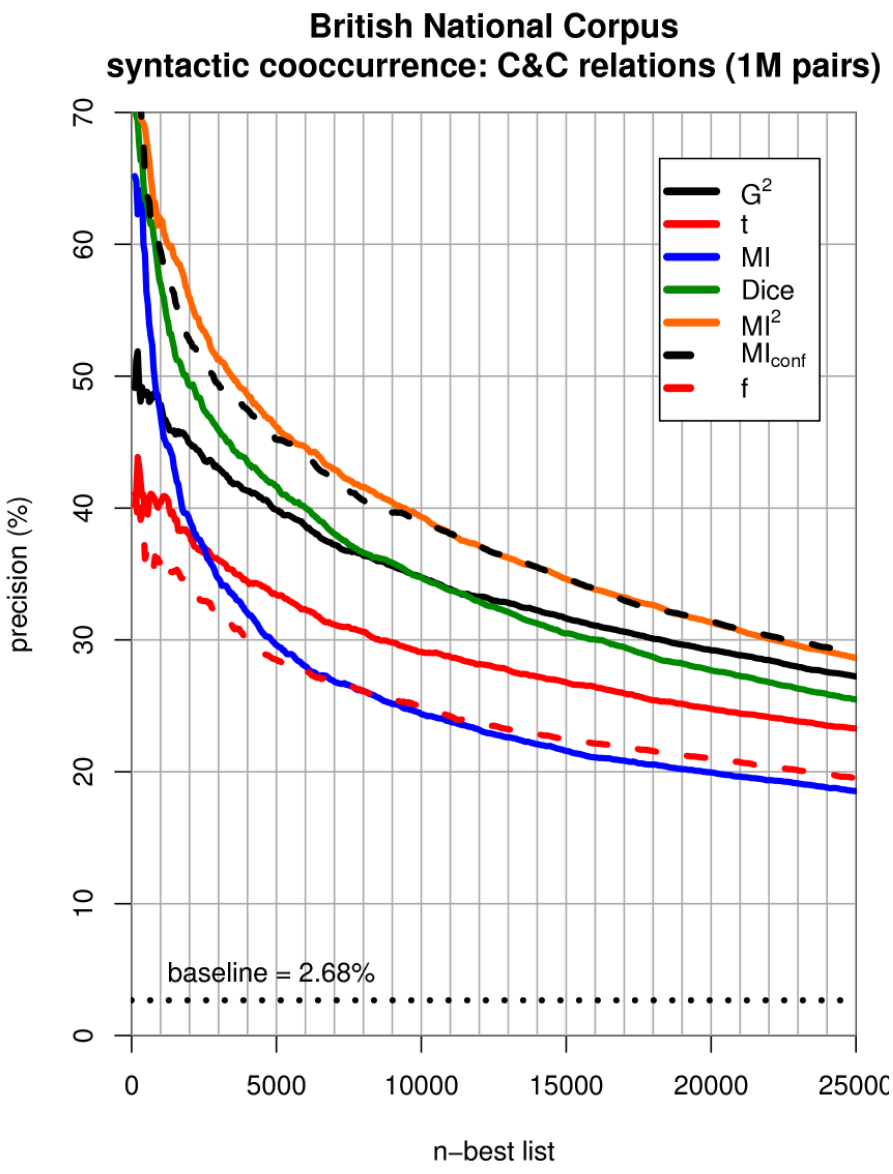


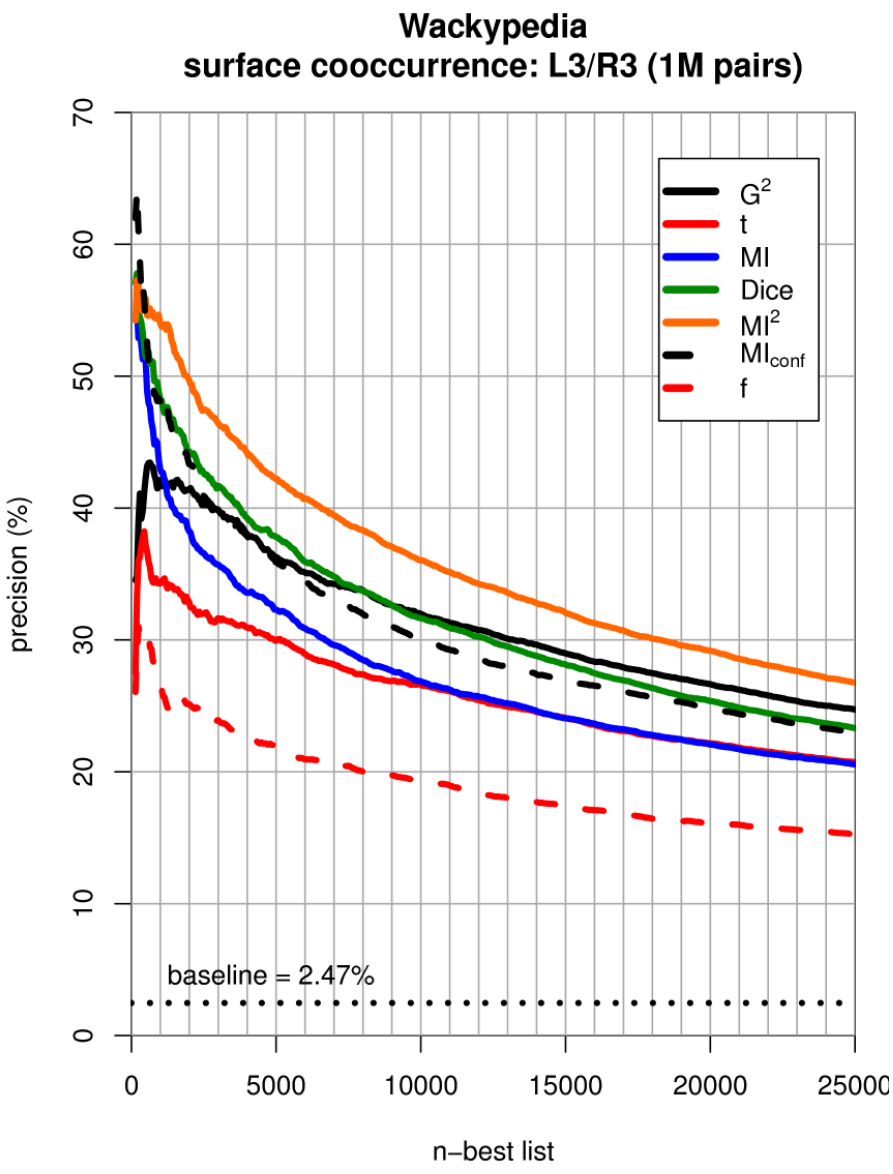
Linguistische Evaluation: BBI-Kollokationen



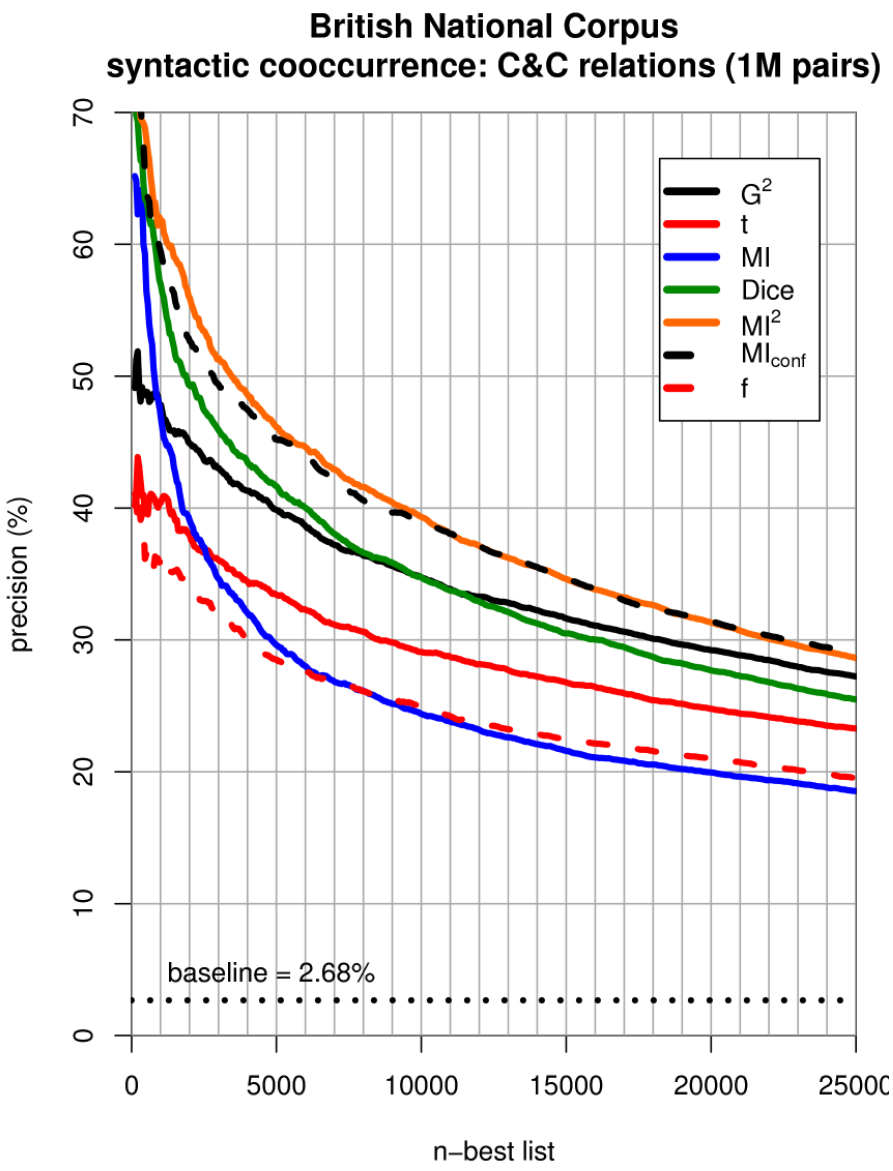


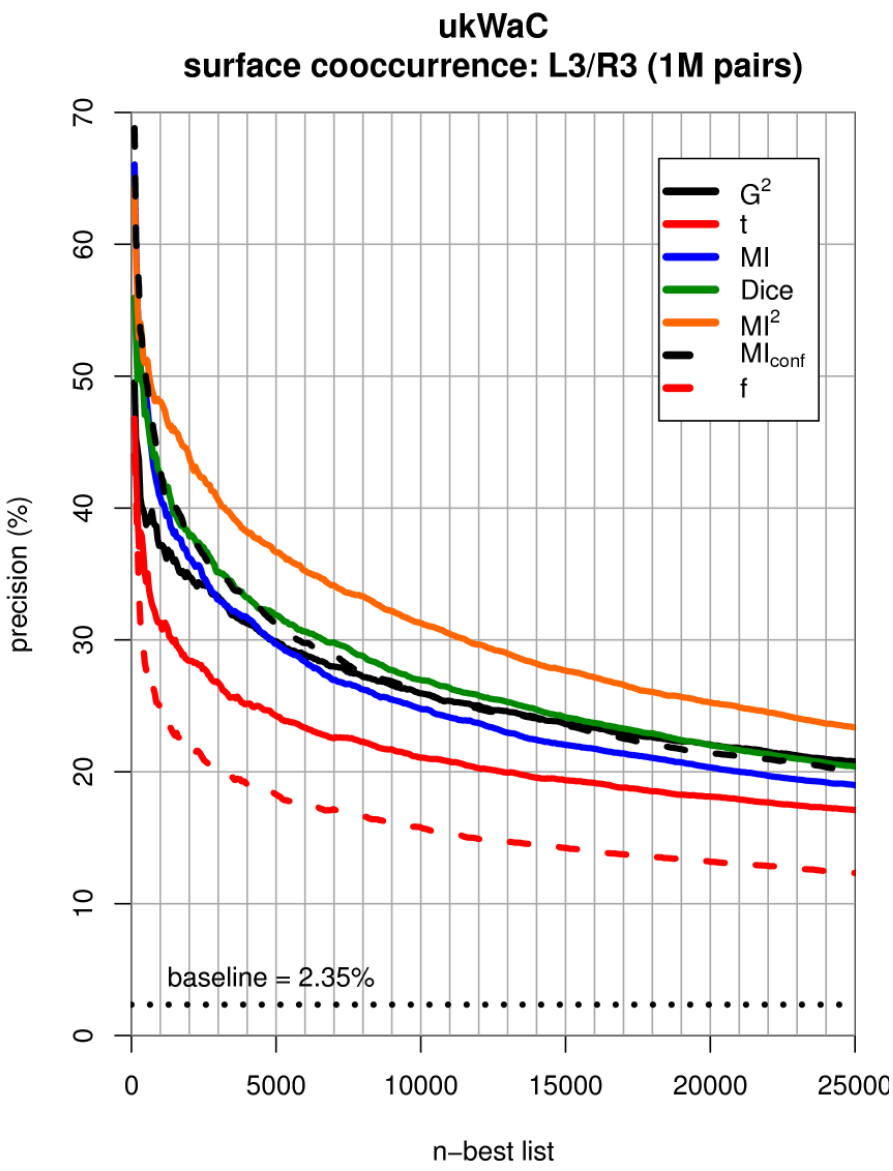
Linguistische Evaluation: BBI-Kollokationen



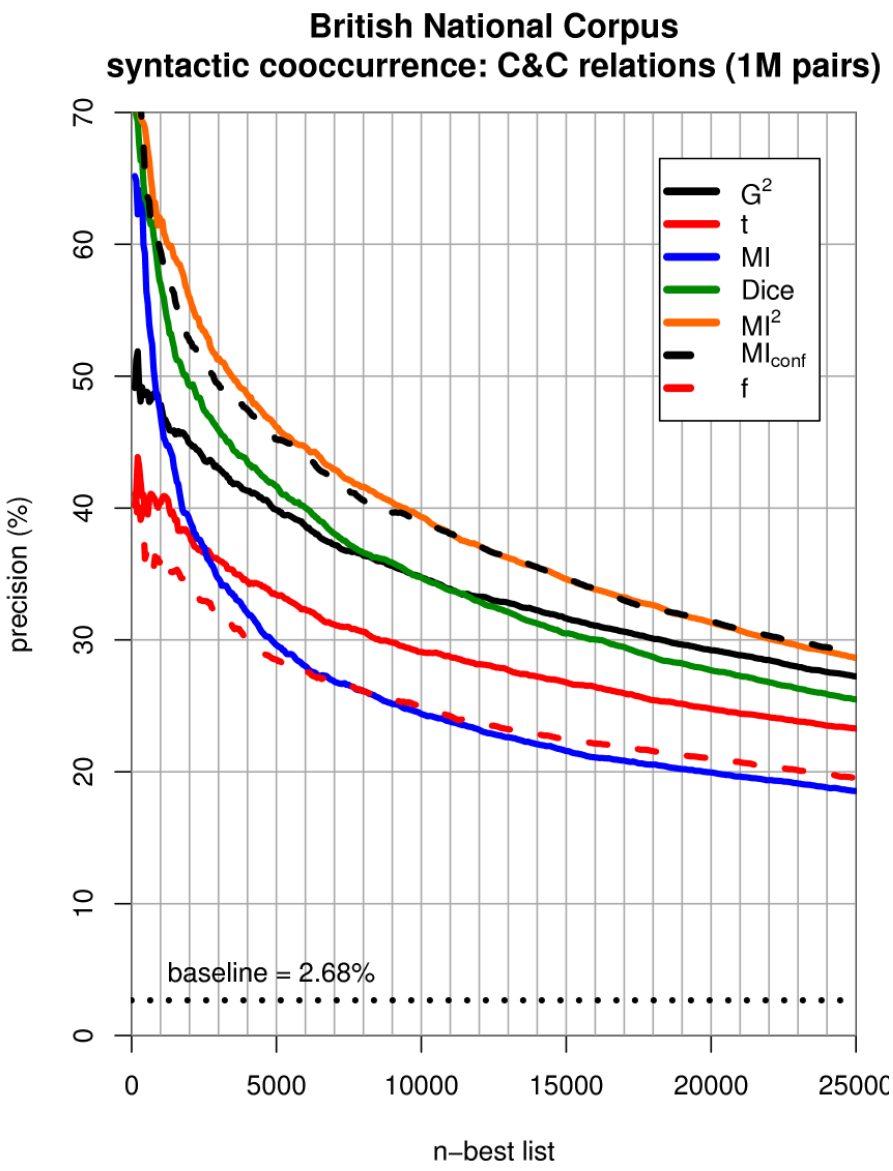


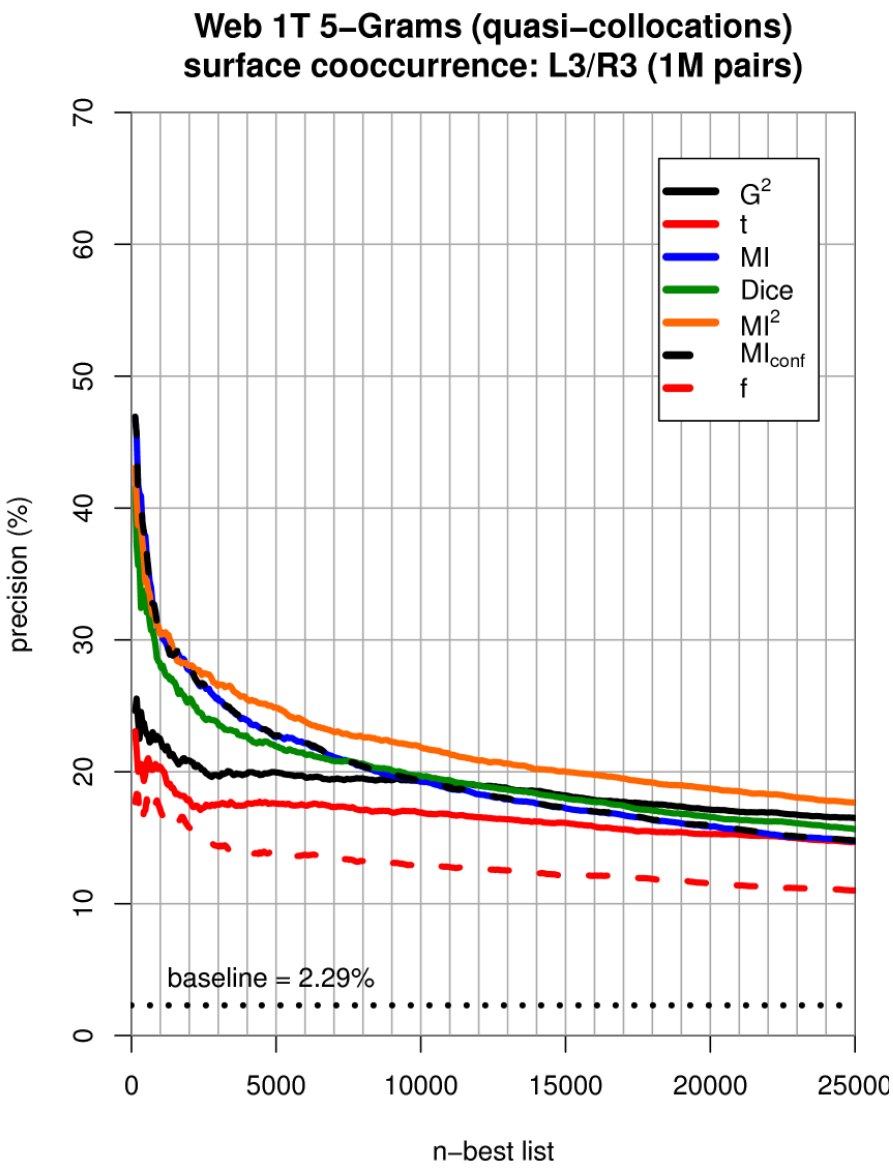
Linguistische Evaluation: BBI-Kollokationen



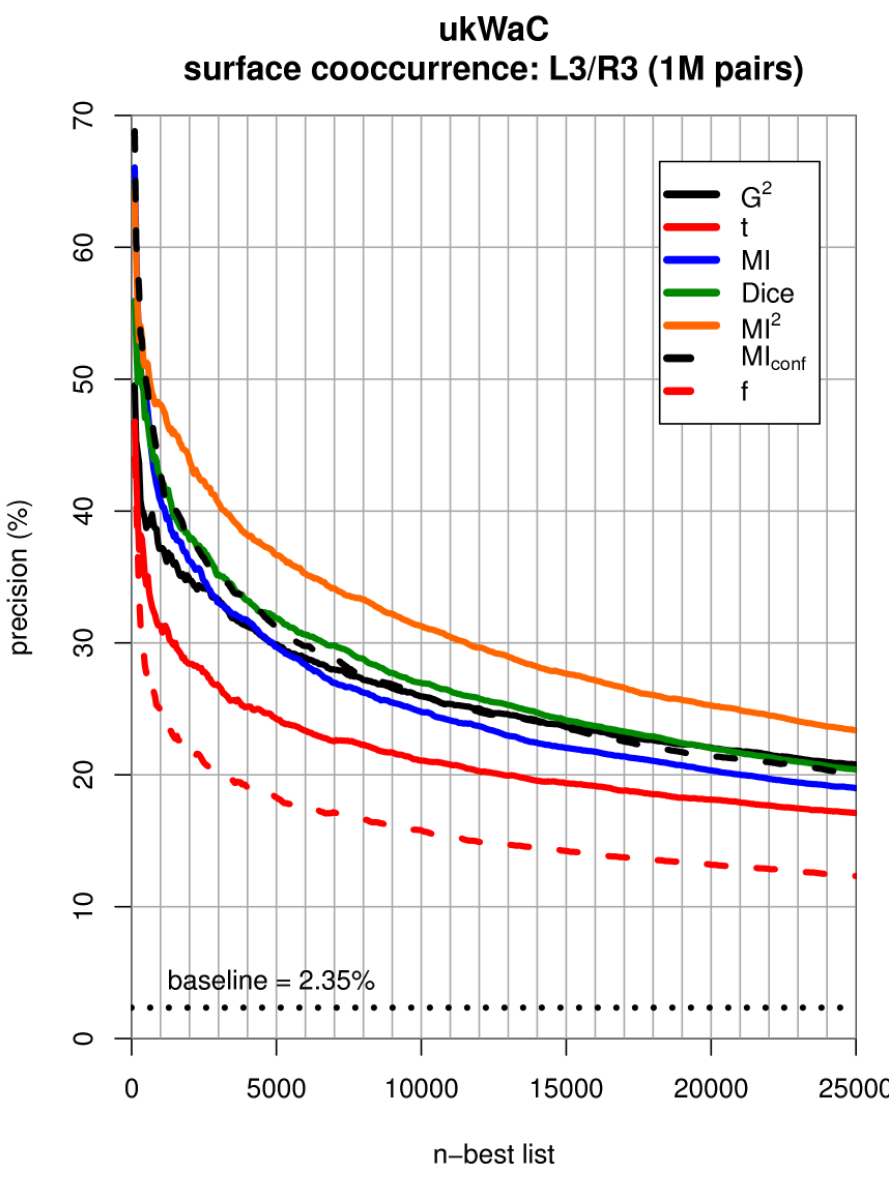


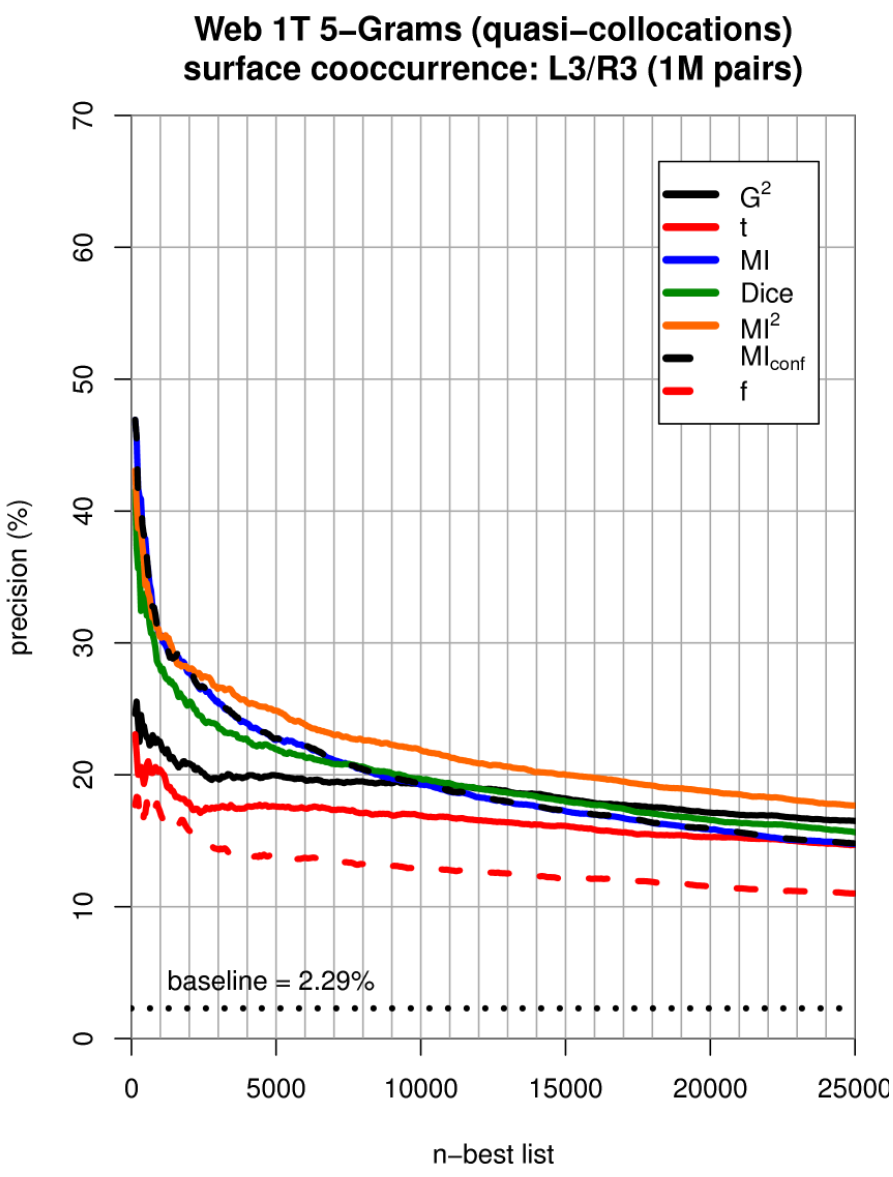
Linguistische Evaluation: BBI-Kollokationen



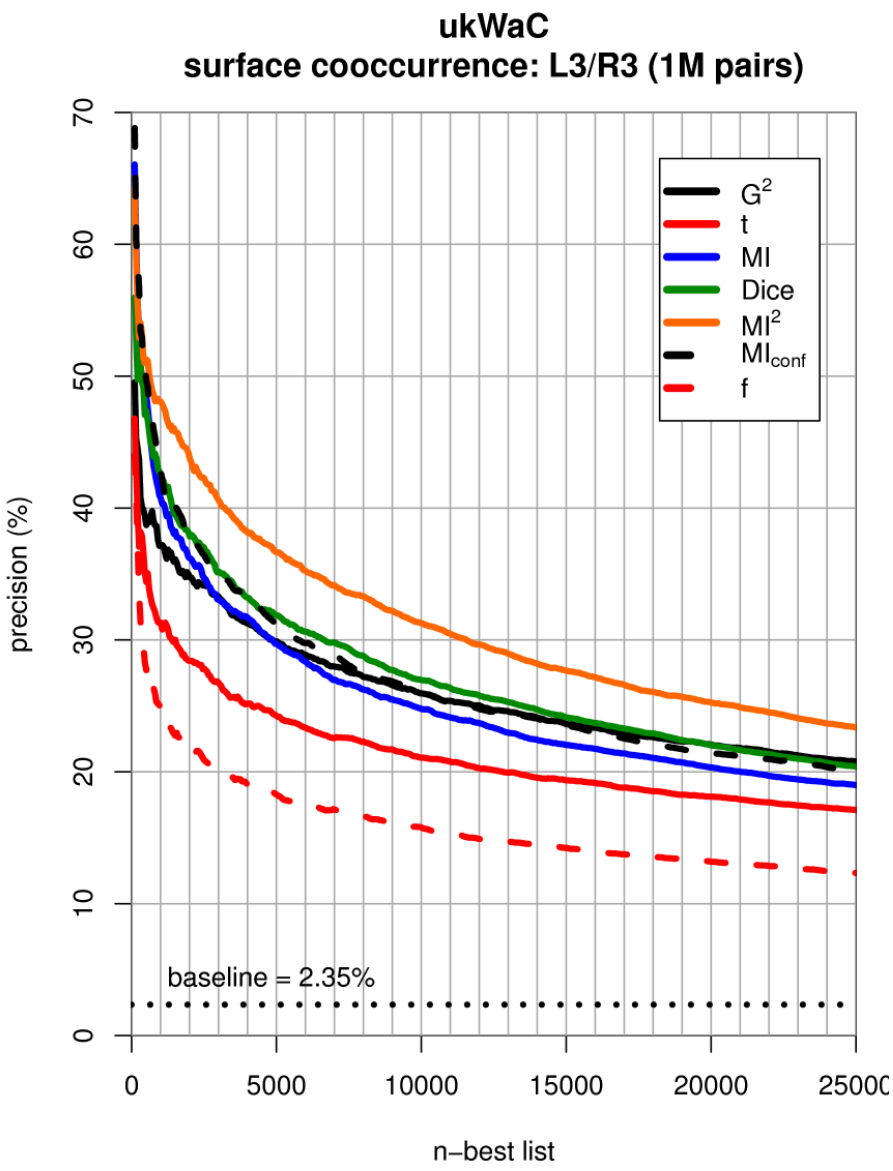


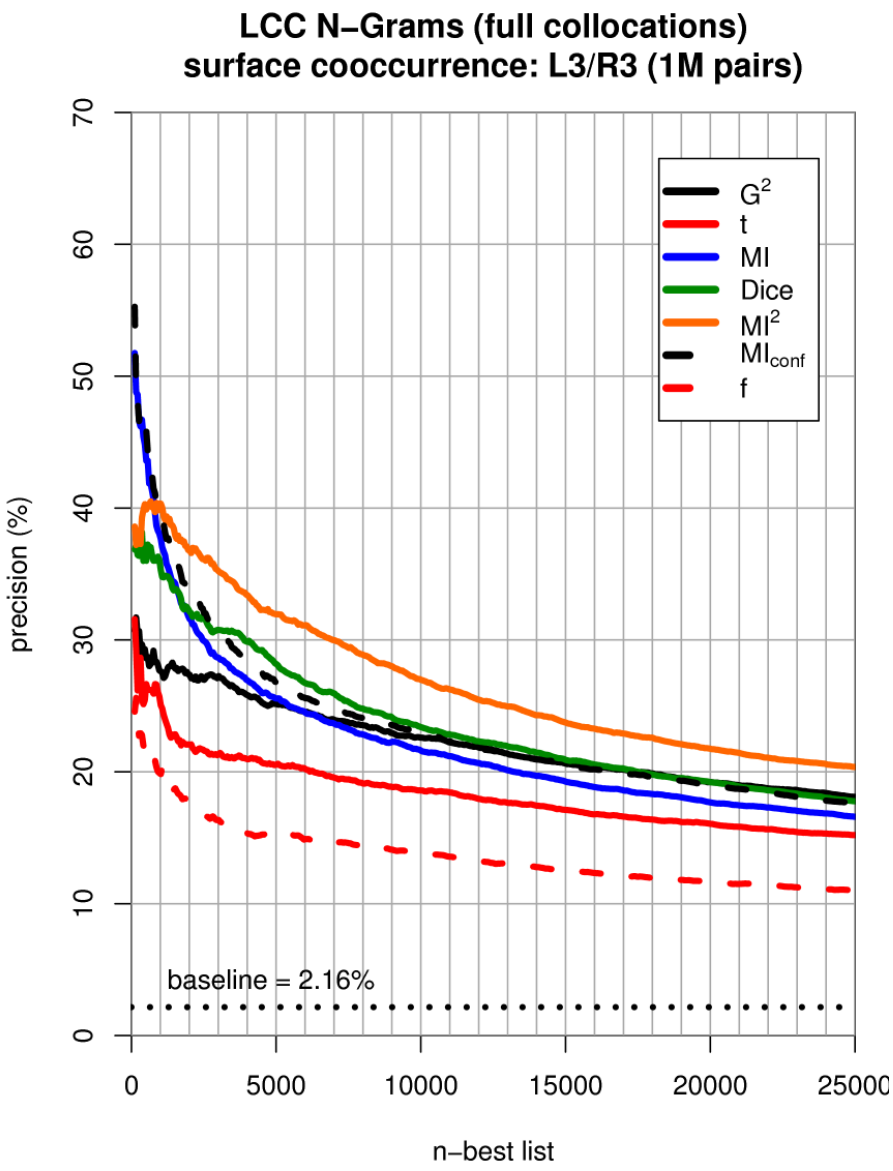
Linguistische Evaluation: BBI-Kollokationen



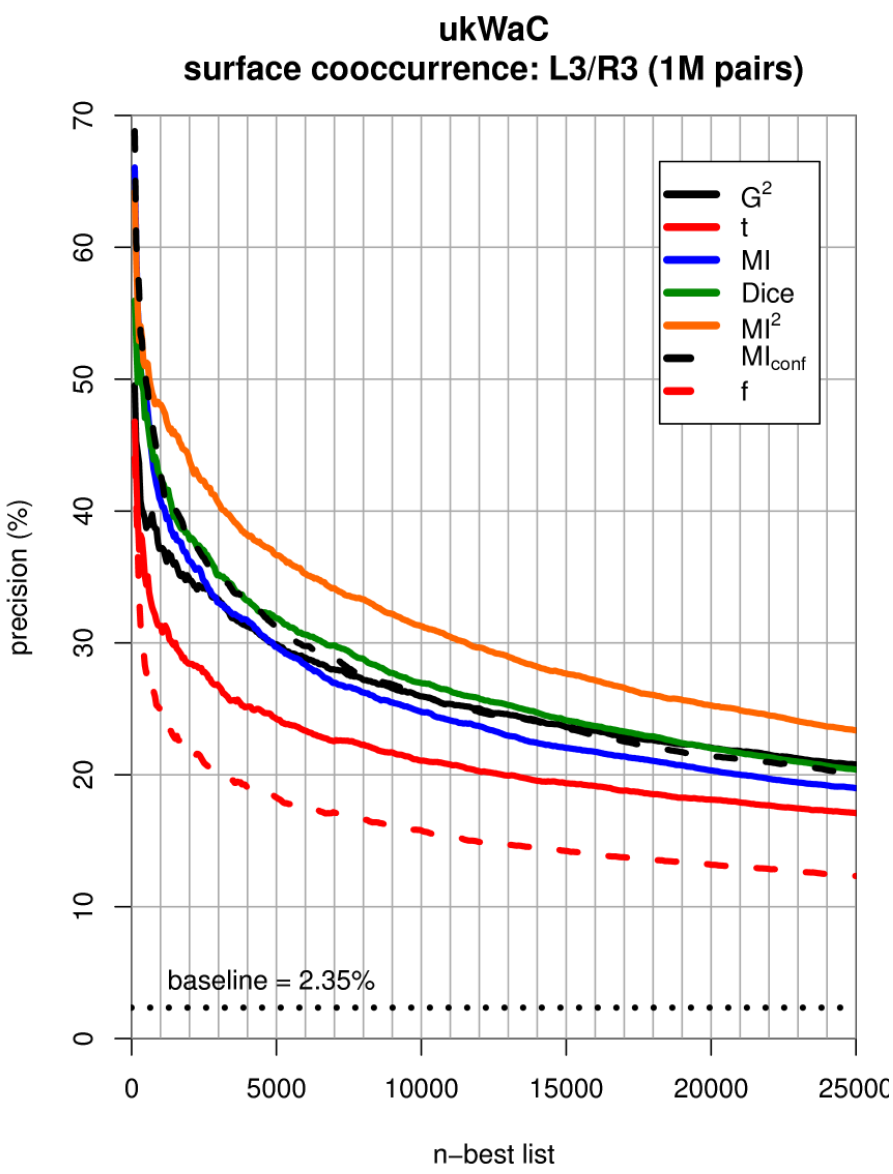


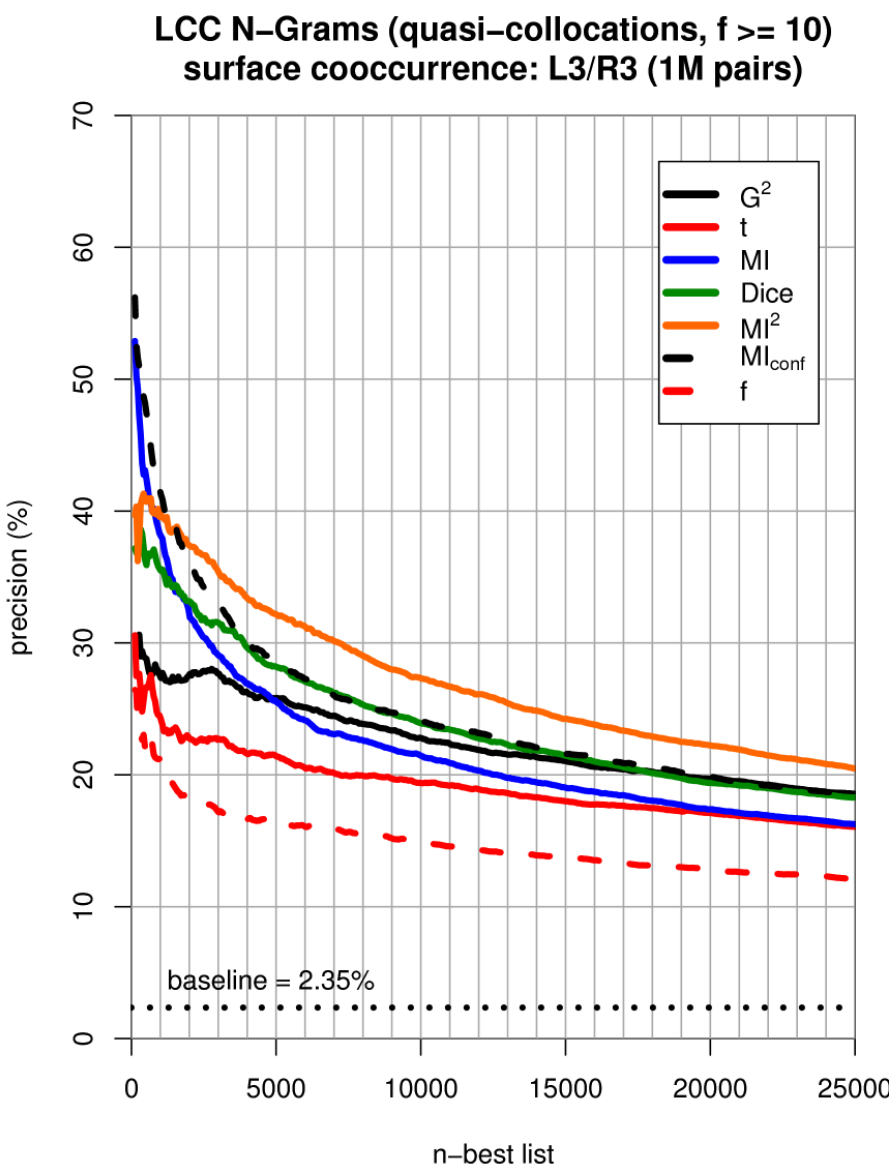
Linguistische Evaluation: BBI-Kollokationen





Linguistische Evaluation: BBI-Kollokationen





DFG Deutsche Forschungsgemeinschaft

THANK YOU!

Evaluation: Distributionelle Semantik

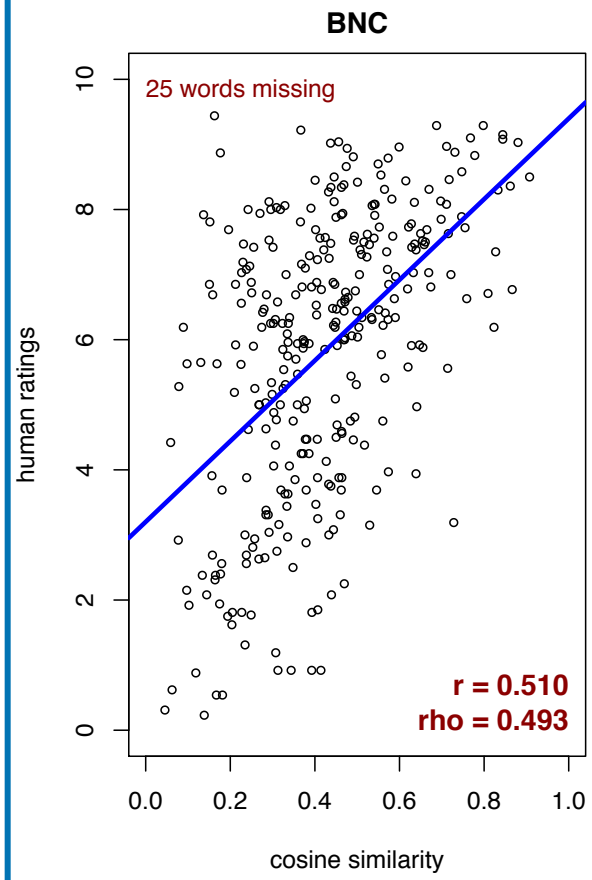
- Wie unterscheiden sich distributionelle Modelle der Wortbedeutung (*wordspace models,* DSM) auf Basis verschiedener Korpora?
  - hier sollte in besonderem Maße *bigger is better* gelten

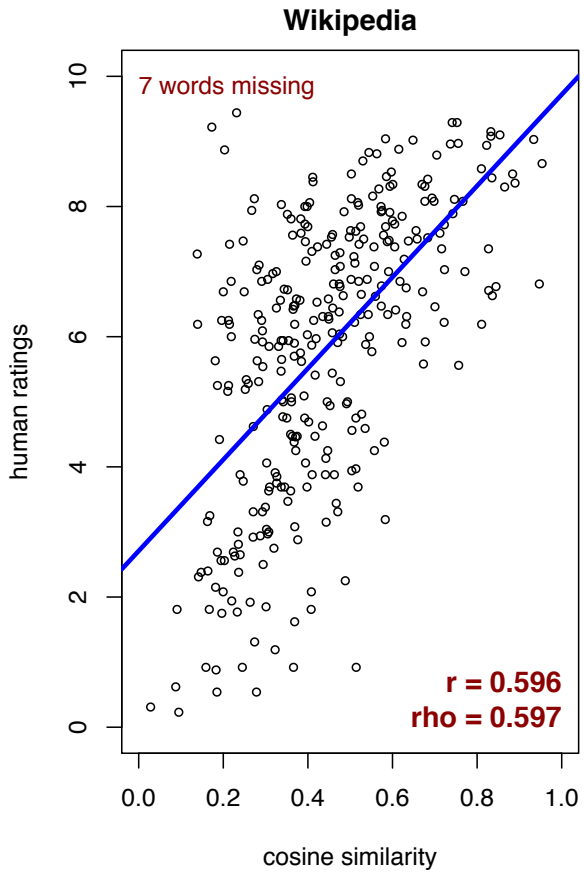
- Evaluation der distributionellen Wortähnlichkeit
  1. Vergleich mit Ähnlichkeitsurteilen durch Versuchspersonen (WordSim-353, Finkelstein *et al.* 2002)
  2. Clustering konkreter Substantive aus 6 semantischen Kategorien (Shared Task bei ESSLLI 2008)

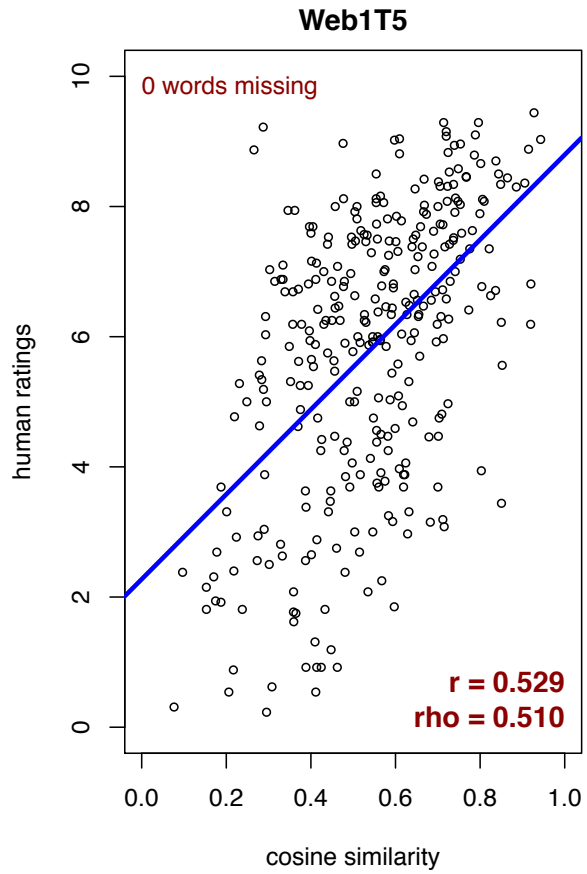
- Basiskorpora: BNC, Wikipedia (WP500), Web1T5, LCC
- Keine Optimierung der DSM-Parameter!
  - term-term matrix, L2/R2 surface context, t-score weighting, sqrt transformation, cosine similarity, SVD to 300 dimensions

Evaluation: Distributionelle Semantik

Korrelation mit WordSim-353

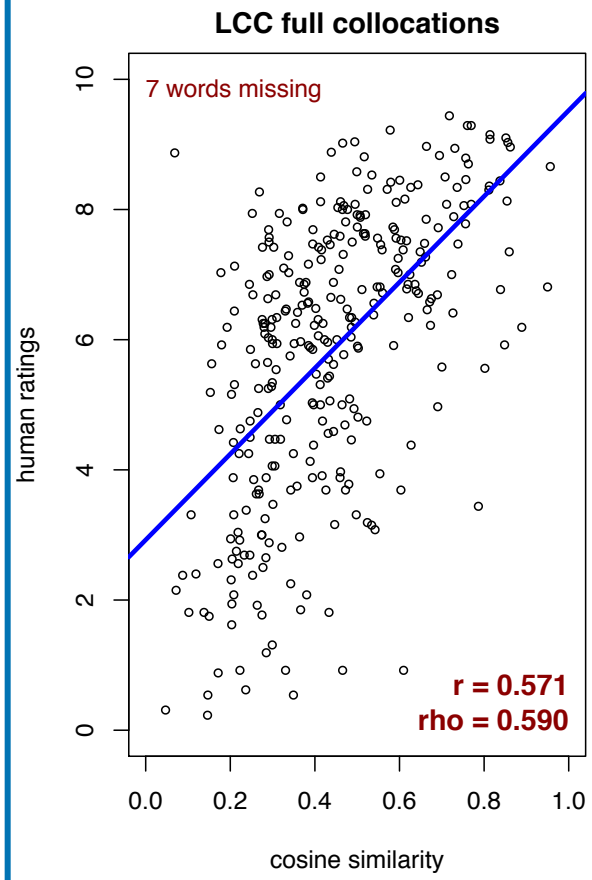


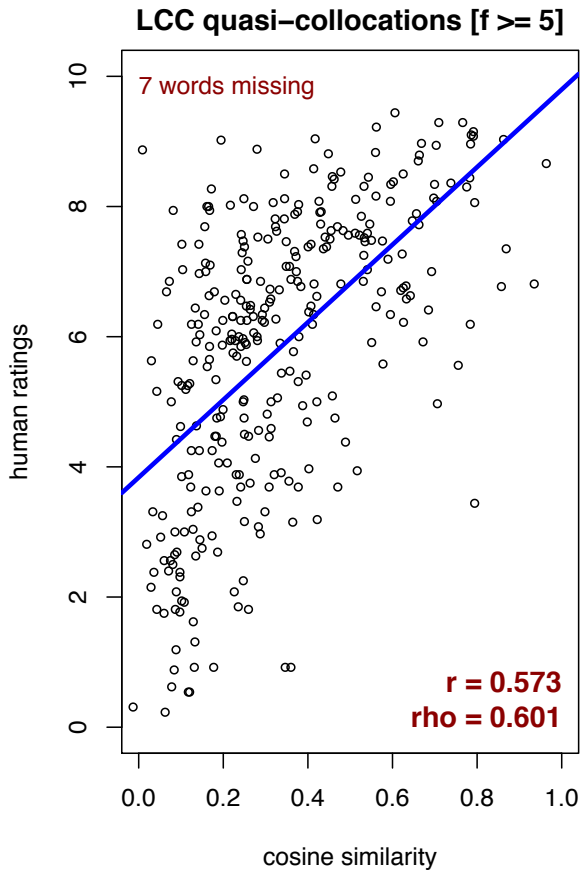


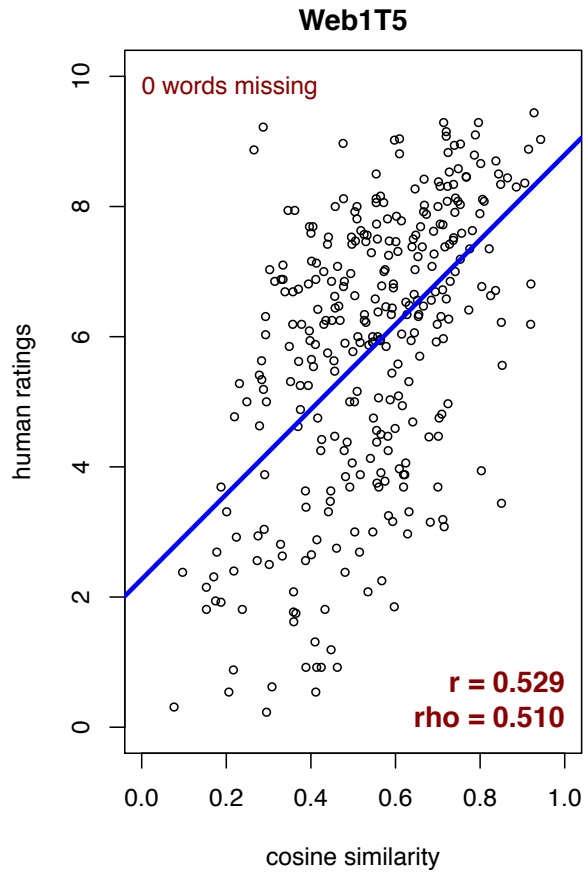


Evaluation: Distributionelle Semantik

Korrelation mit WordSim-353

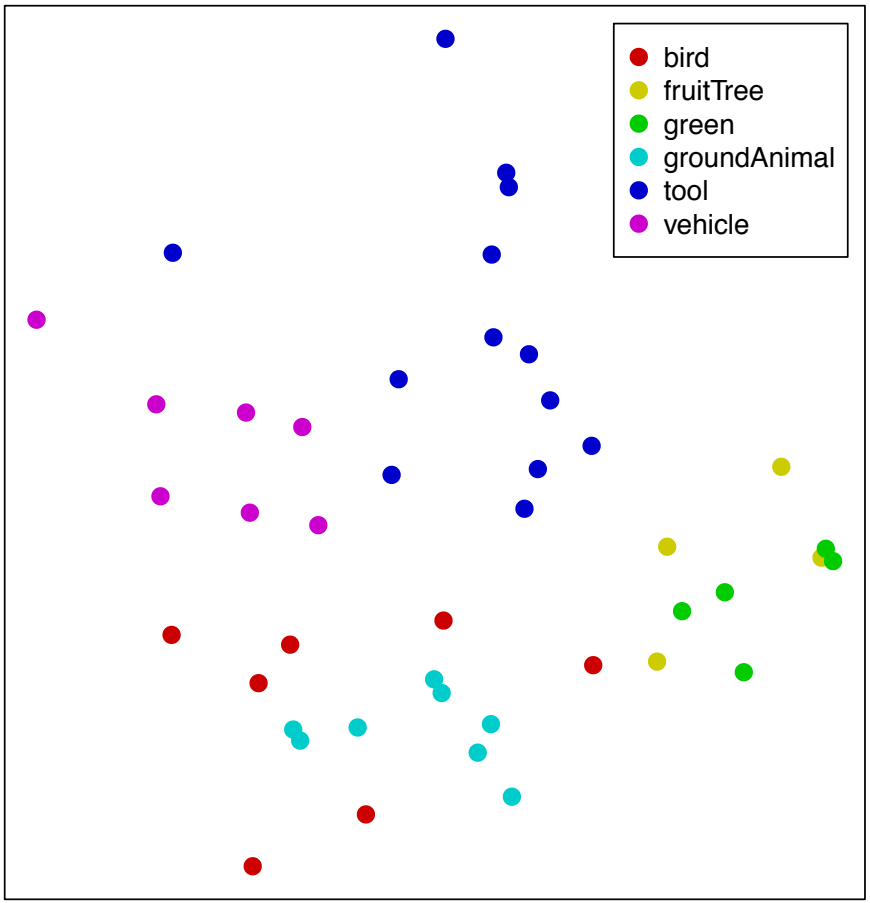


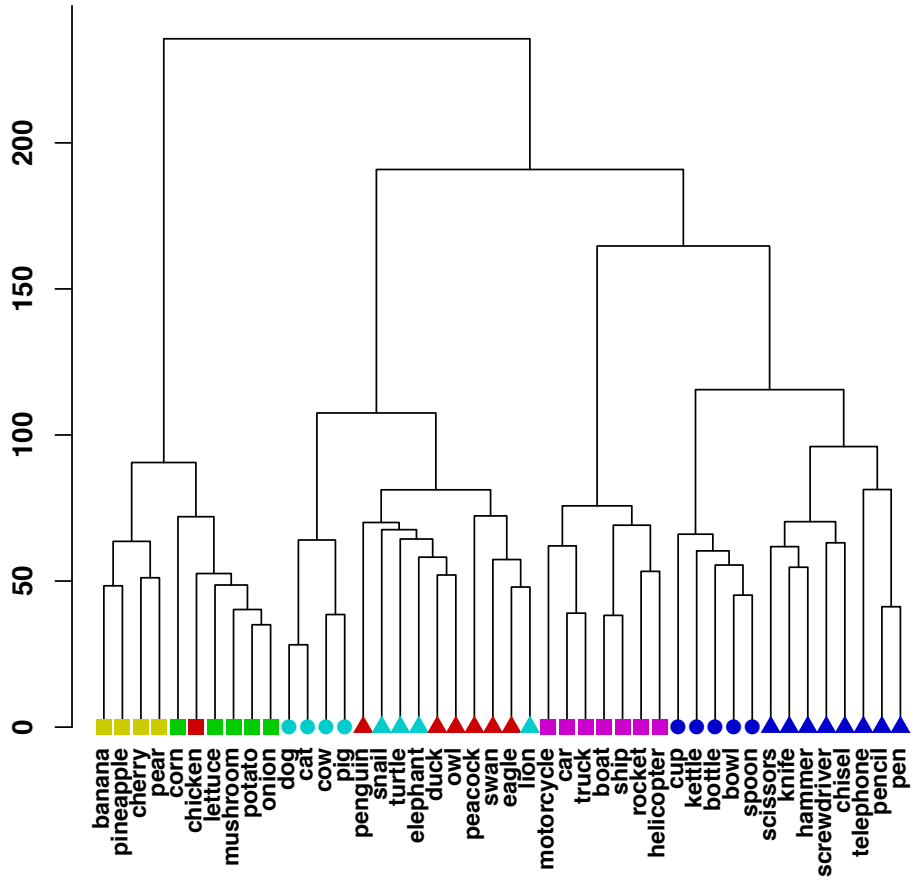




Evaluation: Distributionelle Semantik

ESSLLI Clustering Task





Evaluation: Distributionelle Semantik

ESSLLI Clustering Task

Corpus	purity (hclust)	purity (k-means)
BNC	79.5 %	77.3 %
Wikipedia	79.5 %	77.3 %
Web1T5	75.0 %	77.3 %
LCC	**88.6 %**	**79.5 %**
LCC [f ≥ 5]	**88.6 %**	**79.5 %**

Twitter Example 3

TECHNISCHE UNIVERSITÄT DARMSTADT

@Gunservatively obozo will go nuts when PA elects a Republican Governor next Tue. Can you say redistricting?

@Gunservatively		obozo	will	go	nuts	when	PA	elects	a	Republican	Governor	next	Tue	.	Can	you	say	redistricting	?
@		^	V	V	N	R	^	V	D	A	N	P	^	,	V	O	V	V	,
AT		NP	V	V	NN	ADV	NP	V	ART	ADJ	NN	PP	NP	PUNC	V	PR	V	V	PUNC
@	Gunservatively	obozo	will	go	nuts	when	PA	elects	a	Republican	Governor	next	Tue	.	Can	you	say	redistricting	?
SYM	NP	NP	MD	VV	NNS	WRB	NP	VVZ	DT	NP	NP	JJ	NP	SENT	MD	PP	VVP	VVG	SENT
O	NP	NP	V	V	NN	ADV	NP	V	ART	NP	NP	ADJ	NP	PUNC	V	PR	V	V	PUNC

Twitter Example 3

@Gunservatively obozo will go nuts when PA elects a Republican Governor next Tue. Can you say redistricting?

Specialized Tagset.

@Gunservatively		obozo	will	go	nuts	when	PA	elects	a	Republican	Governor	next	Tue	.	Can	you	say	redistricting	?
@		^	V	V	N	R	^	V	D	A	N	P	^	,	V	O	V	V	,
AT		NP	V	V	NN	ADV	NP	V	ART	ADJ	NN	PP	NP	PUNC	V	PR	V	V	PUNC
@	Gunservatively	obozo	will	go	nuts	when	PA	elects	a	Republican	Governor	next	Tue	.	Can	you	say	redistricting	?
SYM	NP	NP	MD	VV	NNS	WRB	NP	VVZ	DT	NP	NP	JJ	NP	SENT	MD	PP	VVP	VVG	SENT
O	NP	NP	V	V	NN	ADV	NP	V	ART	NP	NP	ADJ	NP	PUNC	V	PR	V	V	PUNC

Twitter Example 3

TECHNISCHE UNIVERSITÄT DARMSTADT

@Gunservatively obozo will go nuts when PA elects a Republican Governor next Tue. Can you say redistricting?

Specialized Tagset.

@Gunservatively		obozo	will	go	nuts	when	PA	elects	a	Republican	Governor	next	Tue	.	Can	you	say	redistricting	?
AT		NP	V	V	NN	ADV	NP	V	ART	**ADJ**	**NN**	**PP**	NP	PUNC	V	PR	V	V	PUNC
@	Gunservatively	obozo	will	go	nuts	when	PA	elects	a	Republican	Governor	next	Tue	.	Can	you	say	redistricting	?
O	NP	NP	V	V	NN	ADV	NP	V	ART	**NP**	**NP**	**ADJ**	NP	PUNC	V	PR	V	V	PUNC

Highlighting high level differences.

UKP

Twitter Example 4

TECHNISCHE UNIVERSITÄT DARMSTADT

lmao s/o to the cool ass asian officer 4 #1 not runnin my license and #2 not takin dru boo to jail . Thank u God . #amen

lmao	s/o			to	the	cool	ass	asian	officer	4	#1	not	runnin	my	license	and	#2	not	takin	dru	boo	to	jail	.	Thank	u	God	.	#amen
!	V			P	D	A	N	A	N	P	$	R	V	D	N	&	$	R	V	N	N	P	N	,	V	O	^	,	#
INT	V			PP	ART	ADJ	NN	ADJ	NN	PP	CARD	ADV	V	ART	NN	CONJ	CARD	ADV	V	NN	NN	PP	NN	PUNC	V	PR	NP	PUNC	HASH
lmao	s	/	o	to	the	cool	ass	asian	officer	4	#1	not	runnin	my	license	and	#2	not	takin	dru	boo	to	jail	.	Thank	u	God	.	#amen
NN	NN	SYM	NN	TO	DT	JJ	NN	NN	NN	CD	NN	RB	NN	PP$	NN	CC	NN	RB	NN	NN	VVP	TO	NN	SENT	VV	NN	NP	SENT	#	NN
NN	NN	O	NN	O	ART	ADJ	NN	NN	NN	CARD	NN	ADV	NN	PR	NN	CONJ	NN	ADV	NN	NN	V	O	NN	PUNC	V	NN	NP	PUNC	O	NN

Twitter Example 4

lmao s/o to the cool ass asian officer 4 #1 not runnin my license and #2 not takin dru boo to jail . Thank u God . #amen

lmao	s/o		to	the	cool	ass	asian	officer	4	#1	not	runnin	my	license	and	#2	not	takin	dru	boo	to	jail	.	Thank	u	God	.	#amen
INT	**V**		**PP**	ART	ADJ	NN	**ADJ**	NN	**PP**	**CARD**	ADV	**V**	**ART**	NN	CONJ	**CARD**	ADV	**V**	NN	**NN**	**PP**	NN	PUNC	V	**PR**	NP	PUNC	**HASH**
lmao	s	/	o	to	the	cool	ass	asian	officer	4	#1	not	runnin	my	license	and	#2	not	takin	dru	boo	to	jail	.	Thank	u	God	.	#amen
NN	**NN**	O	NN	**O**	ART	ADJ	NN	**NN**	NN	**CARD**	**NN**	ADV	**NN**	**PR**	NN	CONJ	**NN**	ADV	**NN**	NN	**V**	**O**	NN	PUNC	V	**NN**	NP	PUNC	O	**NN**

Highlighting high level differences.

Twitter Example 5

RT @torsten_zesch I cannot belive how well this tagger works :) http://code.google.com/p/jwpl/

RT	@torsten_zesch		I	cannot	belive	(	sic	)	how	well	this	tagger	works		:)	http://code.google.com/jwpl/							
~	@		O	V	N	,	R	,	R	R	D	N	V		E	U							
DM	AT		PR	V	NN	PUNC	ADV	PUNC	ADV	ADV	ART	NN	V		EMO	URL							
RT	@	torsten_zesch	I	cannot	belive	(	sic	)	how	well	this	tagger	works	:	)	http	:	/	/	code.google.com	/	jwpl	/
NP	SYM	NN	PP	MD	VV	(	JJ	)	WRB	RB	DT	NN	VVZ	:	)	NN	:	SYM	SYM	NN	SYM	NN	SYM
NP	O	NN	PR	V	V	O	ADJ	O	ADV	ADV	ART	NN	V	O	O	NN	O	O	O	NN	O	NN	O

Twitter Example 5

RT @torsten_zesch I cannot belive how well this tagger works :) http://code.google.com/p/jwpl/

Different Tokenization

RT	@torsten_zesch		I	cannot	belive	(	sic	)	how	well	this	tagger	works		:)	http://code.google.com/jwpl/								
~	@		O	V	N	,	R	,	R	R	D	N	V		E	U								
DM	AT		PR	V	NN	PUNC	ADV	PUNC	ADV	ADV	ART	NN	V		EMO	URL								
RT	@	torsten_zesch	I	cannot	belive	(	sic	)	how	well	this	tagger	works	:	)	http	:	/	/	code.google.com	/	jwpl	/	
NP	SYM	NN	PP	MD	VV	(	JJ	)	WRB	RB	DT	NN	VVZ	:	)	NN	:	SYM	SYM	NN	SYM	NN	SYM	
NP	O	NN	PR	V	V	O	ADJ	O	ADV	ADV	ART	NN	V	O	O	NN	O	O	O	NN	O	NN	O	

Twitter Example 5

RT @torsten_zesch I cannot belive how well this tagger works :) http://code.google.com/p/jwpl/

RT	@torsten_zesch		I	cannot	belive	(	sic	)	how	well	this	tagger	works	:)	http://code.google.com/jwpl/
DM	**AT**		PR	V	**NN**	**PUNC**	**ADV**	**PUNC**	ADV	ADV	ART	NN	V	**EMO**	**URL**
RT	@	torsten_zesch	I	cannot	belive	(	sic	)	how	well	this	tagger	works	:)	http : / / code.google.com / jwpl /
NP	**O**	NN	PR	V	**V**	**O**	**ADJ**	**O**	ADV	ADV	ART	NN	V	**O** **O**	**NN** **O** **O** **O** **NN** **O** **NN** **O**

Highlighting high level differences.