

Übersetzungen im Profil

--

Registershifts und ihre Signifikanz

Silvia Hansen-Schirra

BNC

-

TEC

- <http://www.natcorp.ox.ac.uk/>
- fiction part
- 4,741,500 words
- written by English native speakers

- <http://www.monabaker.com/tsresources/TranslationalEnglishCorpus.htm>
- fiction part
- 4,843,763 words
- translated by professional translators who are English native speakers

BNC

-

TEC

- <http://www.natcorp.ox.ac.uk/>
- fiction part
- 4,741,500 words
- written by English native speakers

- <http://www.monabaker.com/tsresources/TranslationalEnglishCorpus.htm>
- fiction part
- 4,843,763 words
- translated by professional translators who are English native speakers

-
- typical fiction features: past tense verbs, third person pronoun, perfect aspect, public verbs, synthetic negation, present participle clauses, time adverbials, place adverbials, adverbs, phrasal coordination
 - untypical fiction features: present tense verbs, attributive adjectives, *wh*-relative clauses, pied piping, phrasal coordination, nominalizations, conjuncts, agentless passives, *by*-passives, past participle clauses, subordinators, *that*-clauses, demonstratives, final prepositions, existential *there*. (Biber 1995)

BNC

-

TEC

- <http://www.natcorp.ox.ac.uk/>
- fiction part
- 4,741,500 words
- written by English native speakers

- <http://www.monabaker.com/tsresources/TranslationalEnglishCorpus.htm>
- fiction part
- 4,843,763 words
- translated by professional translators who are English native speakers

-
- typical fiction features: past tense verbs, third person pronoun, perfect aspect, public verbs, synthetic negation, present participle clauses, time adverbials, place adverbials, adverbs, phrasal coordination
 - untypical fiction features: present tense verbs, attributive adjectives, *wh*-relative clauses, pied piping, phrasal coordination, nominalizations, conjuncts, agentless passives, *by*-passives, past participle clauses, subordinators, *that*-clauses, demonstratives, final prepositions, existential *there*. (Biber 1995)

-
- specific properties of translations: TL normalization vs. SL shining-through (Teich 2001)
 - universal features of translations: explicitation, simplification, normalization, levelling out (Baker 1996)

Corpus Annotation and Querying

He	PPHS1	and	CC
omitted	WD	that	DD1
to	TO	,	YC
mention	WV0	as	CSA
that	CST	Romulus	NP1
one	MC1	had	VHD
person	NN1	quietly	RR
had	VHD	informed	WN
been	VBN	him	PPH01
killed	WN	,	YC
and	CC	both	DB2
another	DD1	were	VBDR
badly	RR	past	II
wounded	WN	saving	WVG
,	YC	.	VF

TnT tagger:

- statistical part-of-speech tagger (Brants 1999)
- trainable for different languages
- Susanne tagset for English (Sampson 1995) & Stuttgart-Tübingen tagset (STTS) for German (Schiller et al. 1999)
- tokenization tool
- between 30,000 and 60,000 tokens per second
- accuracy of about 97 %
- output: tab separated vector (TSV)

Corpus Annotation and Querying

He	PPHS1	and	CC
omitted	WVD	that	DD1
to	TO	,	YC
mention	WV0	as	CSA
that	CST	Romulus	NP1
one	MC1	had	VHD
person	NN1	quietly	RR
had	VHD	informed	WVN
been	VBN	him	PPH01
killed	WVN	,	YC
and	CC	both	DB2
another	DD1	were	VBDR
badly	RR	past	II
wounded	WVN	saving	WVG
,	YC	.	VF

TnT tagger:

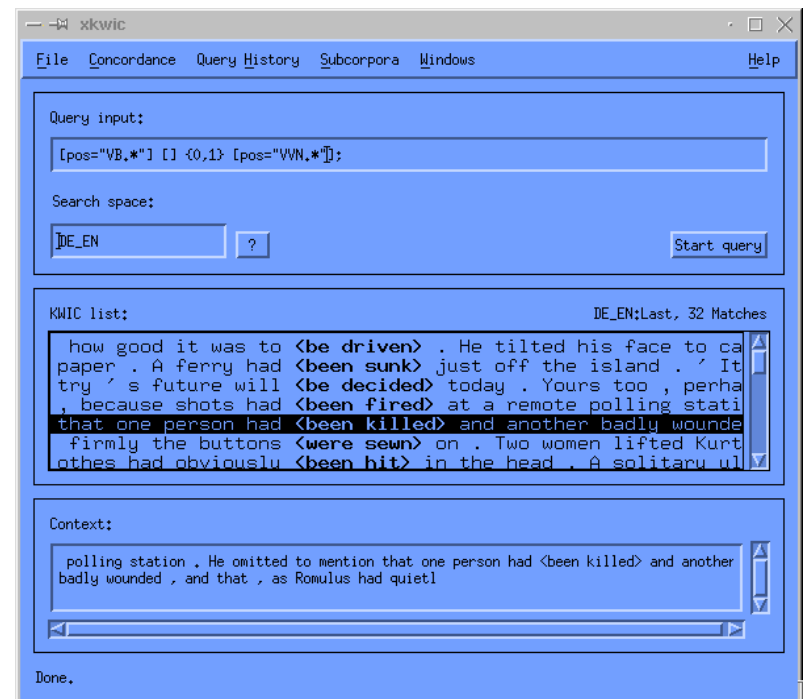
- statistical part-of-speech tagger (Brants 1999)
- trainable for different languages
- Susanne tagset for English (Sampson 1995) & Stuttgart-Tübingen tagset (STTS) for German (Schiller et al. 1999)
- tokenization tool
- between 30,000 and 60,000 tokens per second
- accuracy of about 97 %
- output: tab separated vector (TSV)

IMS Corpus Workbench:

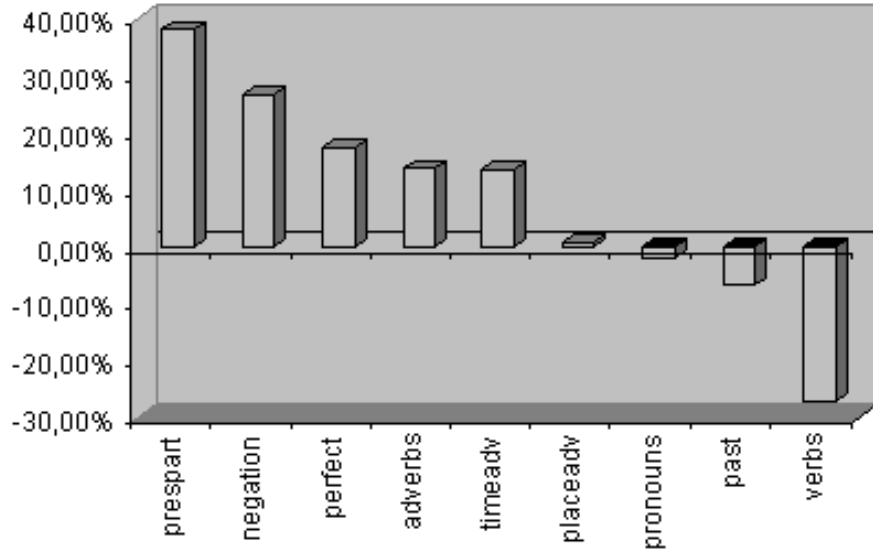
- query for words and/or part-of-speech tags on the basis of regular expressions (Christ 1994)
- two modules: CQP (corpus query processor) and Xkwic (user interface)
- KWIC (keyword in context) list
- sample query: part-of-speech tags VB.* (forms of the verb *be*) followed by VVN.* (past participle) and zero to one words in between

Significance testing:

- norming + chi square test (Oakes 1998)



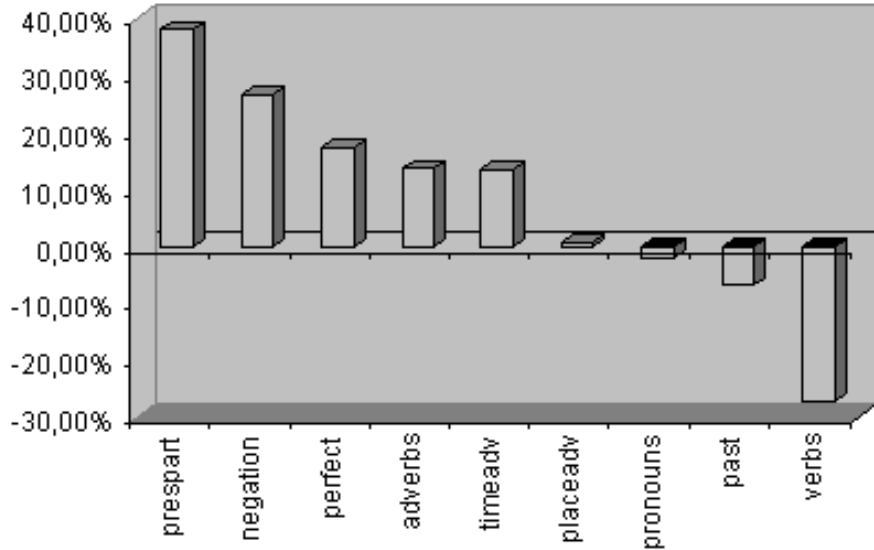
BNC vs. TEC: Register Shift



Typical fiction features:

- higher frequency of typical fiction features in TEC
- overuse of typical fiction features in translations
- translators conform to the typical patterns of the target language
- the use thereof being even exaggerated
- translations are even more typical of their register
- trend towards **normalization** for typical fiction features

BNC vs. TEC: Register Shift

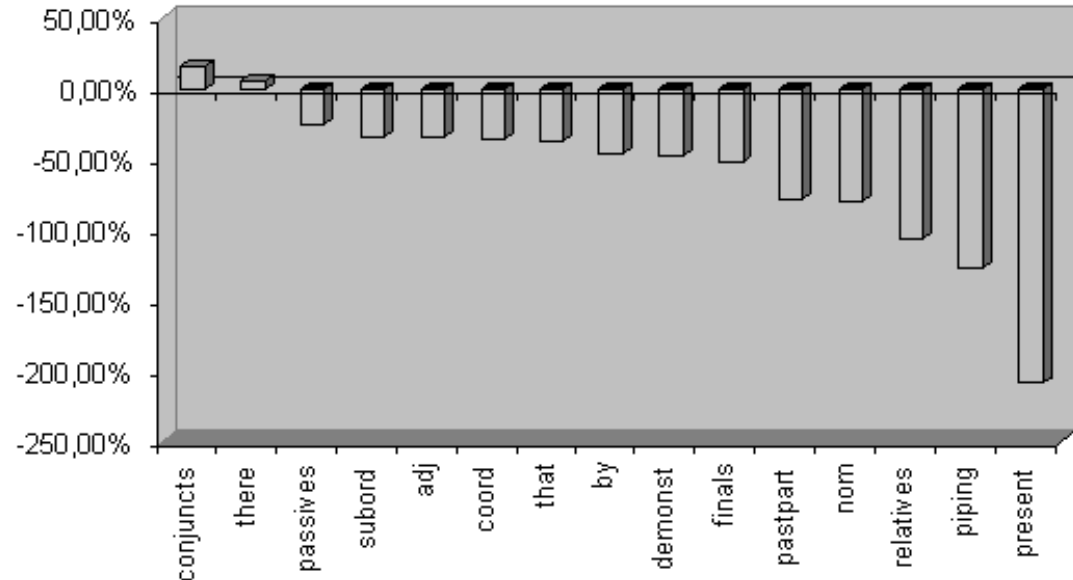


Typical fiction features:

- higher frequency of typical fiction features in TEC
- overuse of typical fiction features in translations
- translators conform to the typical patterns of the target language
- the use thereof being even exaggerated
- translations are even more typical of their register
- trend towards **normalization** for typical fiction features

Untypical fiction features:

- higher frequency of untypical fiction features in TEC
- overuse of untypical fiction features in translations
- translators do not conform to the typical patterns of the target language
- translations represent a more neutral register
- trend towards **anti-normalization** for untypical fiction features



Vergleich mit CroCo (English Fiction, typische)

	TEC	BNC	CroCo- Ori	CroCo-Trans
Adverben	6,13 %	5,39 %	5,94 %	7,32 %
Pronomen 3. Person	4,93 %	5,03 %	5,25 %	4,85 %
Verben in der Vergangen- heitsform	3,85 %	4,11 %	4,21 %	3,70 %
Verben im Perfekt	0,89 %	0,76 %	0,72 %	0,69 %
Partizip I	0,83 %	0,60 %	0,30 %	0,28 %
Verben des Redens	0,58 %	0,74 %	0,65 %	0,54 %

Vergleich mit CroCo (English Fiction, typische)

	TEC	BNC	CroCo- Ori	CroCo-Trans
Adverben	6,13 %	5,39 % ↓	↓ 5,94 %	7,32 %
Pronomen 3. Person	4,93 %	5,03 % ↑	↑ 5,25 %	4,85 %
Verben in der Vergangen- heitsform	3,85 %	4,11 % ↑	↑ 4,21 %	3,70 %
Verben im Perfekt	0,89 %	0,76 % ↓	↑ 0,72 %	0,69 %
Verben des Redens	0,58 %	0,74 % ↑	↑ 0,65 %	0,54 %

Offene Fragen für die statistische Auswertung

- Problem 1: unterschiedliche Grundmengen
- Problem 2: unterschiedliche Datenbasis
- Problem 3: kleine Zahlen
- Wie erkennt man Korrelationen bzw. Musterbildung?
- Wie können abhängige und unabhängige Variablen miteinander in Bezug gesetzt werden, um sprach- und texttypologisch bedingte Veränderungen zu klassifizieren?