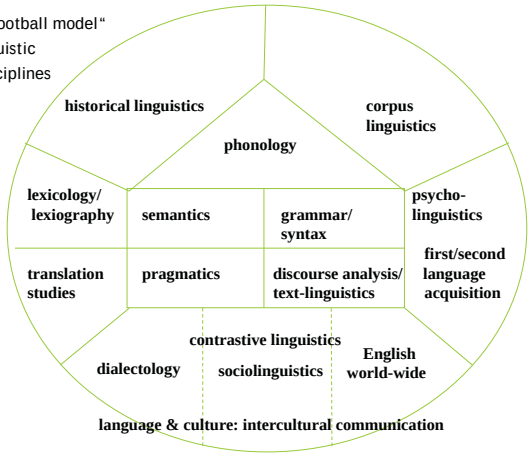
School of Foreign Languages, Sun Yat-sen University,
April 15, 2015

An Introduction to Corpus-Linguistics:
Principles & Applications, from Teaching to Research?

Josef Schmied
English Language & Linguistics
Chemnitz University of Technology
https://www.tu-chemnitz.de/phil/english/sections/ling/presentations_js.php
josef.schmied@phil.tu-chemnitz.de

The „football model“ of linguistic subdisciplines



Survey of Applied Linguistics

1. Text linguistics/Discourse Analysis
2. Sociolinguistics
3. Psycholinguistics
4. Second-Language Acquisition (SLA) / ELT
5. Corpus-Linguistics
6. Lexicography
7. Translation Studies
8. Language & Culture/Politeness/Intercultural Communication (ICC)

→ AL, as a “hyphenated” discipline:
partly methods from sociology, psychology, anthropology, etc.
not included here: independent disciplines like

- clinical linguistics
- computational linguistics
- forensic linguistics (authorship: background of asylum seekers, plagiarism, etc.)
  = “to see more than meets the eye”

5. Corpus Linguistics (CL)

5.1 Definition

corpus1 = body or collection of written or spoken material
upon which linguistic analysis is based (structuralism)

corpus2 = machine-readable
“representative”, i.e. stratified “model”
i.e. more than a text collection!

for computer-based language analysis:
corpus-informed (language awareness/ELT)
corpus-based vs. -driven (research)

5.1.2 Tools corpus-analysis software:

WordSmith (with ICAME CD)
Sara (with BNC in TUC Bib)
AntConc downloadable free from
http://www.antlab.sci.waseda.ac.jp/antconc_index.html

reality in photo and abstraction

5.1.3 CL: methodology or theory?

- pro methodology:
  - CL is not parallel to phonology, syntax, lexicology or pragmatics (core linguistics)
  - CL is not restricted to any linguistic level (can be used to address phonological, syntactic, pragmatic etc. questions, [as is sociolinguistics → ?])
  - “corpus” no reference to area of linguistic investigation
    (vs. sociolinguistics, psycholinguistics, computational linguistics, etc.)
  - methodologies adopted from social sciences inform sociolinguistics, but are not a theory in themselves (participant observation, interviews, etc.)
- pro theory:
  - CL has a particular outlook on language
    - rules of language are usage-based, not normative (as in prescriptive grammars)
    - linguistic change occurs when speakers use L for communication
  - CL introduced new methods and principles which have theoretical status ≠ theory

→ pro methodology e.g. combined with SFL

5.1.4 The corpus-based vs. the intuition-based approach

- descriptions of English strongly biased (personal views)
- normative description do not take variation into consideration
- authenticity of invented examples can be questionable
- introspection-based results not verifiable,
  e.g. "I get myself a soft drink." (*pop*, *soda*, *coke* considered incorrect)
- personal opinion is reflecting idiolect, not real speech
- professor's shoeboxes
  - Jespersen kept thousands of notes of real English from literary texts
  - first to use them as authentic examples in his grammar
- corpus data can address preferences / tendencies; quantitative questions in general (frequency effects etc.)
- improved reliability of corpus-based over intuition-based approach
- BUT: corpus-based approach not suitable for all research questions → approaches should be seen as complementary rather than exclusive

5.1.5 Reasons

for the popularity of corpus linguistics, esp. among non-native speakers

because it combines a qualitative and quantitative perspective

- offers citations used as real language samples of language usage
- provides a view beyond individual experience
- rules out individual salience
- computer processable

output: - concordances (KWIC=key word in context)
- collocates (*milk gets sour*, *butter rancid*, *eggs addled* = groceries spoil differently)
- relative vs. absolute frequencies --> "normalise" = per 1 M. words

AntConc example

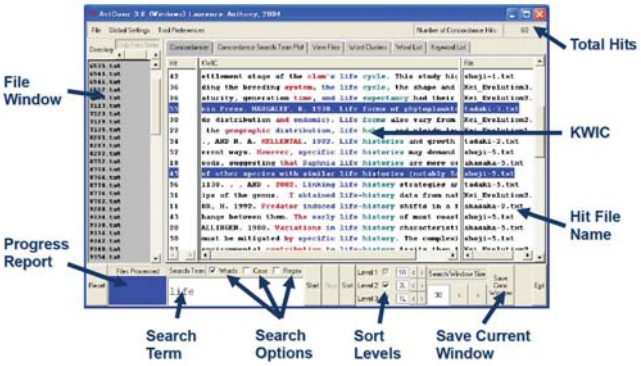


Figure 1 KWIC Concordancer Tool

5.2 Corpus-based vs. corpus-driven approaches

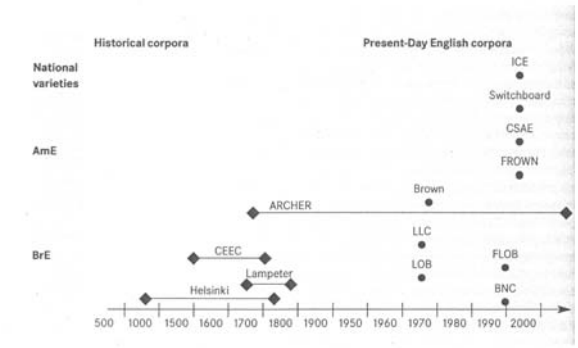
1) type of corpus data
   - representativeness (c-driven: large corpora balance themselves; c-based: belief in natural balance unwarranted)
   - corpus size (c-driven: the larger the better; BUT: generally only analyse every nth instance – unclear how this is different from c-based methodology)
2) attitudes towards existing theories and intuitions
   - corpus annotation (c-driven: strong objections)
   - c-driven: *tabula rasa* ideal → no preconceived idea as to results; BUT: in reality, c-driven approaches make use of traditional categories such as word classes, etc. without defining them (which is a de-facto annotation …)
   - c-based: typically start out from a theoretical issue / problem and use corpus data to illustrate / solve it
3) research focus
   - c-driven: holistic approach to language description (since such notions as pragmatics or syntax demand a theory)
   - c-based: focus on individual levels of linguistics
4) corpus-based approach by no means as radical as corpus-driven approach
   - c-driven approach "claims to be a new paradigm within which a whole language can be described"

5.3 Developments in corpus compilation
5.3.1 50 years of corpus history

(forerunners 1950s American structuralists, e.g. Harris)
1959 Quirk: Survey of English Usage (SEU)
1,000,000 words written/spoken 1953-1987
>London-Lund corpus of spoken English
1963/64 Francis/Kucera: Brown Corpus
1M of written American English from 1961
1970-1978 Johansson & Leech: LOB parallel to Brown
1M written BritE (Lancaster-Oslo/Bergen Corpus) from 1961
1980 - Cobuild Corpus (Birmingham, Sinclair) -> Bank of English
1990 - International Corpus of English (ICE):
UK, US?, CA, AU/NZ, EA (KE/TZ), ZA, HK, SG, IN, PH, etc.
1990 - International Corpus of Learner English (ICLE)
1990 - 1993 British National Corpus 100M (10M spoken)
from 1990 Freiburg Corpora: FLOB and Frown from 1991 etc. (parallel LOB/Brown)
for recent language change

since 1998 www as "corpus" (WebCorp, WebPhraseCount)

5.3.2 Corpora on the history and variation of English



Kortmann (2005: 36)

5.3.3 Reference corpora on the WWW

corpus.byu.edu

corpora, size, queries = better resources, more insight

Created by Mark Davies, BYU. Overview, search types, researchers, publications, corpus-based resources.

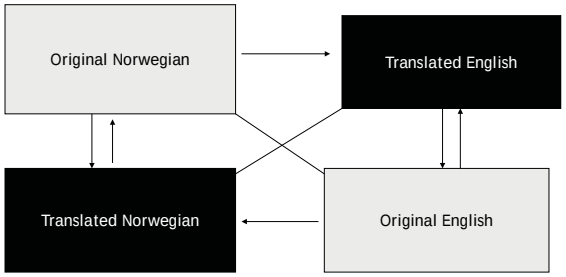
English	# words	language/dialect	time period	compare
Global Web-Based English (GloWbE)	1.9 billion	20 countries	2012-13	
Corpus of Contemporary American English (COCA)	450 million	American	1990-2012	* * * * *
Corpus of Historical American English (COHA)	400 million	American	1810-2009	* *
TIME Magazine Corpus	100 million	American	1923-2006	
Corpus of American Soap Operas	100 million	American	2001-2012	*
British National Corpus (BYU-BNC)*	100 million	British	1980s-1993	* *
Strathy Corpus (Canada)	50 million	Canadian	1970s-2000s	

Academic English Corpora:

- Michigan Corpus of Academic Spoken English (MICASE)
- Michigan Corpus of Upper-Level Student Papers (MICUSP)
- British Academic Spoken / Written English (BASE / BAWE)

cf. ChemCorpus of student / academic writing (theses)

Corpus design translation / parallel corpora



Johansson, S & Hofland, K (1993) *Towards an English-Norwegian Parallel Corpus*

5.4 Corpus compilation principles

5.4.1 Corpus types

large and stratified:

- mega-/reference corpora
  - British National Corpus (BNC)
    90 M written/10 M spoken, demographic/context-governed from 1991-94
    http://www.natcorp.ox.ac.uk/
  - American / Australian National Corpus being compiled now (problematic; ANC 20 M)
- 'national corpora', e.g. ICE http://www.ucl.ac.uk/english-usage/ice/
  e.g. ICE-East Africa, ICE-Canada (parallel corpora)
- genre/domain specific corpora
  e.g. SPACE (Specialised & Popular ACademic English),
  Trains (dialogue corpus)
- translation corpora
  e.g. EU corpus
  English-German Translation Corpus
- 'quick and dirty'/ad-hoc corpora, e.g. for translation problems, not translating

5.4.2 Representativeness / Balance in corpus design

Leech: representative = findings can be generalised (What)
Biber: representative findings show the same degree and extent of variability as the total population (How)

a representative corpus must contain different text types / genres / registers
possibility of constructing a monitor corpus, depending on the view on a corpus (dynamic vs. static)

in a general corpus, balance and sampling are responsible for achieving representativeness
acceptable level of balance also depends on intended uses (e.g. specialized corpora)

N.B.: "any claim of corpus balance is largely an act of faith rather than a statement of fact as, at present, there is no reliable scientific measure of corpus balance" (McEnery, Xiao & Tono 2006: 16)

researchers often adopt earlier corpus construction procedures (primarily BNC)
→ National Corpora in Australia, US, etc.

5.4.3 Sampling

sample = scaled-down version of a larger population
sampling units: e.g. a book, newspaper, periodical
sampling frame: the list of sampling units actually used in the corpus compilation (e.g. all books available in one particular library)
target population: group to be represented in the corpus
sampling techniques:

- simple random sampling: sampling units are numbered, elements are chosen based on a list of random numbers; problem: rare types / genres may not be selected
- stratified random sampling: divides population into groups (*strata*), samples each stratum at random

sample size:

- use full texts or text chunks from written sources?
- if chunks, where from (initial, middle, end chunks)?

→ again, these should be balanced (no either-or)
proportion & number of samples from each category

5.4.4 Annotation / mark-up

Types of mark-up:

- structural markup: descriptive information about the texts
- "metalanguage" → structure of electronic documents (e.g. structure of conversations, categorizing parts of speech, segmenting of spoken or written text, marking of overlapping speech)
- bibliographic information about written text (genre, number of words, tagger which assign part-of-speech), ethnographic information about individuals in spoken texts (e.g. age, gender, social class, region; usually very limited)
- part of speech markup: part-of-speech designation (e.g. noun, verb); produced by software program called *tagger* (e.g. *CLAWS*; 95% accuracy)
- grammatical markup: parses grammatical structures (e.g. phrases, clauses); produced by software program called *parser* (usually 70 – 80% accuracy)
- always manual checks necessary

5.4.5 Annotation procedures

taggers
- assign part-of-speech designations to each word in a sentence
- first tagging programme 1971 by Greene and Rubin
- out of this programme developing of CLAWS tagset by University of Lancaster (still widely used in its updated form)
- according to Leech (1997: 25-6) tagsets should strive for:
  - conciseness: tags should be as short as possible
  - perspicuity: tags should be as readable as possible
  - analysability: tags should have order and hierarchy above more specific tags

taggers are of two types:
- rule-based: based on rules of grammar written into the tagger → e.g. EngCG-2
- probabilistic: based on statistical likelihood that a given tag will occur in a given context; can be trained on corpora

→ the larger the tagset, the greater the accuracy of tagging

5.5 Corpus search strategies

5.5.1 Pattern types: investigating context

collocation = the appearance of one particular word form in certain distance of another particular word forms
> different meanings can have different collocates

colligation = the appearance of one particular word form in a particular grammatical structure

connotation = the semantic environment, can have positive or negative value ("semantic prosody")
e.g. *happen, cause, attempt, try, fail*

collostruction analysis has 3 methods (Wikipedia):
- collexeme analysis, to measure the degree of attraction/repulsion of a lemma to a slot in one particular construction;
- distinctive collexeme analysis, to measure the preference of a lemma to one particular construction over another, functionally similar construction; multiple distinctive collexeme analysis extends this approach to more than two alternative constructions;
- covarying collexeme analysis, to measure the degree of attraction of lemmas in one slot of a construction to lemmas in another slot of the same construction.

http://en.wikipedia.org/wiki/Collostructional_analysis (15/12/13)

5.5.2 Types of frequency: exploring vocabulary

absolute vs. relative frequency of a word form
(standard) deviation from mean frequency of word forms

5.5.3 Corpus research examples

- How frequent is a particular morphological form/grammatical structure?
- Which particular structures have particular meanings?
- Which particular structures have particular locations in texts?

corpus tasks have degrees of complexity

- relevance of tagging:
  - parts-of-speech (POS), e.g. CLAWS tagging for LOB (http://ucrel.lancs.ac.uk/claws/trial.html)
  - semantic: semantic web/web 3.0 (http://en.wikipedia.org/wiki/Semantic_Web)

5.6 Corpus applications

5.6.1 Computational linguistics
5.6.2 Lexicography
5.6.3 Academic Writing

- Corpus compilation
  variables: genres/text types, (sub-)discipline, gender, L1/MT, …
- Corpus analysis --> interpretation
  - research hypotheses confirmed/refuted
  - research hypotheses developed

5.6.1 Towards computational linguistics

= an interdisciplinary field dealing with the statistical and/or rule-based modelling of natural language from a computational perspective
- not limited to any particular field of linguistics
- traditionally, performed by computer scientists who had specialized in the application of computers to the processing of a natural language

= often grouped under artificial intelligence today, but that has older applications (1950s) as well:
- language analysis: tagging, parsing, annotation (5.6.1)
- machine translation: SYSTRAN (5.6.2)
- text processing: spell checkers, style checkers, automatic text production (abstraction/summarisation), Q&A systems
- speech recognition and synthesis (telephone/communication systems)
- others: expert/dialogue systems, CALL, etc.

Language analysis:
part-of-speech tagging + syntactic annotation/treebanks + semantic web

EXAMPLE OF POS TAGGING from LOB (CLAWS1 tagset):

hospitality_NN is_BEZ an_AT excellent_JJ virtue_NN ,_, but_CC not_XNOT when_WRB the_ATI guests_NNS have_HV to_TO sleep_VB in_IN rows_NNS in_IN the_ATI cellar_NN !_! the_ATI lovers_NNS ,_, whose_WP$ chief_JJB scene_NN was_BEDZ cut_VBN at_IN the_ATI last_AP moment_NN ,_, had_HVD comparatively_RB little_AP to_TO sing_VB '_' he_PP3A stole_VBD my_PP$ wallet_NN !_! '_' roared_VBD Rollinson_NP ._.

square brackets (sequential/horizontal, no indentation; Lancaster):
[S[N Nemo_NP1 ._, [N the_AT killer_NN1 whale_NN1 N] ._, [Fr[N who_PNQS N][V 'd_VHD grown_VVN [J too_RG big_JJ [P for_IF [N his_APP$ pool_NN1 [P on_II [N Clacton_NP1 Pier_NNL1 N]P]N]P]J]V]Fr]N] ,_, [V has_VHZ arrived_VVN safely_RR [P at_II [N his_APP$ new_JJ home_NN1 [P in_II [N Windsor_NP1 [safari_NN1 park_NNL1]N]P]N]P]V] ._. S]

from: http://www.comp.lancs.ac.uk/computing/research/ucrel/annotation.html#treebank

semantic tagging → semantic web:
Tim Berners-Lee has described the semantic web as a component of "Web 3.0"
http://en.wikipedia.org/wiki/Semantic_Web

5.6.2 The corpus revolution in lexicography: word-watching → corpus compilation/analysis

today all dictionaries are based on large-scale corpora, esp. the BNC

- new lexical entries are found
- existing lexical entries are enriched by additional information extracted via corpus analysis (e.g. most common forms, connotation, etc.)
- important aspects of word meaning and grammar are highlighted, which were simply never noticed by linguists who had no data to work with
- word frequency analysis is used for annotating lexical entries
- collocational information is collected, organized, and presented (e.g. idiom identification)
- (domain specific) knowledge is extracted
- lexical items unlikely to be found in dictionary sources are extracted (e.g. proper nouns)
- real examples showing how central and typical features of English are used are provided
- paradigmatic- and syntagmatic-driven semantic clustering is performed

5.6.2 Collocations in the Dictionary

A collocation

- is a sequence of words which co-occur more often than would be expected by chance
- refers to the restrictions on how words can be used together, e.g. prepositions are with verbs, or verbs with nouns
- not be confused with idioms (=fixed syntagmatic combinations)

from http//:wasps.itri.brighton.ac.uk

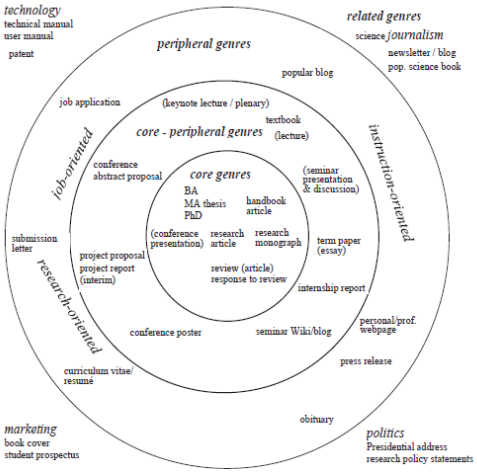
alcohol (as modifier)	BNC freq.	MI score (=mutual information)
alcohol consumption	131	34.0
alcohol abuse	114	31.3
alcohol intake	53	18.2
alcohol misuse	23	17.7
alcohol content	35	15.3
alcohol problem	38	11.3
alcohol dependency	5	10.1
alcohol dependence	7	9.2

since *misuse* is less frequently used than *content*, MI is higher although the absolute frequency is lower

5.6.3 Academic Writing

Matrix of genre types in Academic Writing

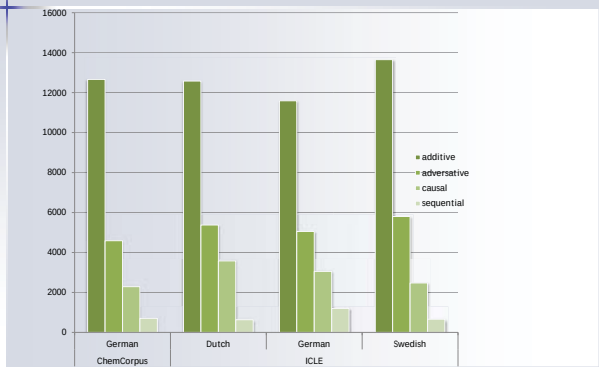
CONTEXT	GENRE	type	medium	audience/ readership	expertise in years	length	purpose	comments
article	**research** (scholarly) **article**	r	w	specialist	5	5,000+	drive research	in double-blind peer-reviewed journal with impact factor
book	research book	r	w	specialist	6	10,000	drive research	also monograph, trend towards IMRAD
	textbook	r	w	novice	8	6-10,000	introduce novices	also course book, introduction?
	handbook	r	w	specialist	10+	10,000+	state-of-the-art	honour=reliable?
	book reviews	r	w	specialist	4	1,000	evaluate/critique	also section in book/thesis
	state-of-the-art review	r	w	specialist	10+	2,000	start project?	research field survey
	article collection (ed)	r	w	specialist	10	150+p	state-of-the-art?	multiple authors,1 topic/area;conference?
project doc.	project proposals	r	w	specialist?		5,000	accepted for qualif./funding	a promise with milestones to check progress
	BA/MA project proposal	r	w	learner	3	1,000	qualification	trend to IMRAD, research questions/hypotheses
	PhD project application	r	w	learner	5	5,000	qualification/funding	trend to IMRAD, research questions/hypotheses
	(thesis/PhD) defense	r	s	learner	3-5y	20'-20-40m	qualification	trend to IMRAD, research questions/hypotheses; onen discussion
	project/progress report	r	w	specialist		6-10,000	justify expense	multiple authors,1 topic/area
conference	conference presentations	r	s	specialist	5	15'-5m	drive research?	protect area? ask advice?
	key-note (lecture)	r	s	specialist	10+	45-60+15-30m	state-of-the-art?	famous scholar,personality
	plenary (lecture)	r	s	specialist	10+	30-60m+15	research overview	broad topic?
	progress report	r	s	specialist	5	10m	demonstrate research? ask advice?	
	(conference) abstract, proposal	r	s	specialist	5	300-600w	acceptance	
	poster introduction	r	s	specialist	5	3m	view poster	very comics=attractive?
	conference poster	r	w	specialist	5	1 A0	introduce research,popularise?	Rivend to IMRAD; fig/tables=ref.s
	conference proceedings (ed)	r	w	specialist	10+	150+p	document research	multiple authors,1 topic/area
	conference report	r?	w	public	3	1,000	drive research?	popular? dissemination
university	lecture	t	s	learner	10+	45-90m	disseminate knowledge	competes with textbook? problem attention span
teaching	student/seminar presentation	t	s	learner	1	10-20m	qualification	problem discussion? self-protective?
	Wiki	t/r?	w	learner	1		collaborate in Knowledge creation	groupswork/multiple authors
	classroom discussion	t		learner	1		collaborate in knowledge creation	problem interaction
	field notes	t	w	self?	1		collect information	data collection methodologies?
	BA thesis	t/r?	w	specialist	3	40+n	qualification	
	MA thesis	r	w	specialist	5	60+p	qualification	
	PhD thesis	r	w	specialist	8	200+p	qualification	conventions,not a research book
	habilitation/postdoctoral thesis	r	w	specialist	10+	200+p	qualification	tend to small, reslicted by articles in p-r i.?
subsidiary?	(article) abstract	r	w	specialist	5	1-300w	read? full article	part (free; to decide worth paying?)
	handout	t/r?	w	Ext	1	1-2n	support,take-home	large diagrams,figures,statistics; examples;references
"valorisation"	university journal/newsletter	r	w	public	5	1-2n	demonstrate "value"	untrained?
	popular blog	r	w	public?	1	1+1+1	time-line of development (projects,career)	technical platform for old stagers?
	popular science book	t/r?	w	public	8	80-200?	create interest in research?	



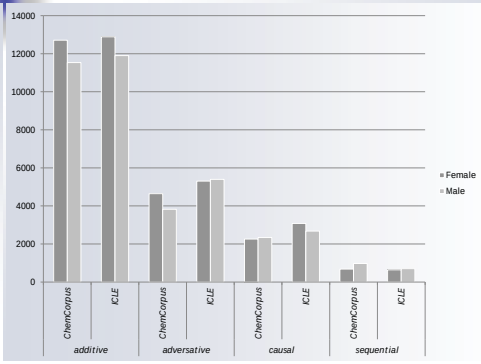
ChemCorpus set-up by genre, degree programme and specialisation

genre	specialisation	number of texts	average length	total words
timed Mag paper	language/linguistics	70	4,200	0.3 Mill.
Magister thesis	language/linguistics	25	25,000	0.6 Mill.
	culture/literature	11	30,000	0.4 Mill.
Total Magister		106		1.3 Mill.
term paper BA	language/linguistics	100	4,200	0.5 Mill.
	culture/literature	100	4,700	0.5 Mill.
project report	(cultural)	120	4,000	0.5 Mill.
BA thesis	language/linguistics	80	12,000	1 Mill.
	culture/literature	80	16,000	1 Mill.
term paper MA	language/linguistics	80	5,700	0.5 Mill.
	culture/literature	80	6,600	0.5 Mill.
MA thesis	language/linguistics	40	25,000	1 Mill.
	culture/literature	40	25,000	1 Mill.
Total B/MA		720		6.5 Mill.

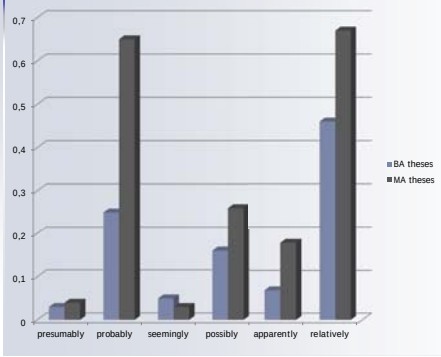
Functional linkers by L1 (per one million words)
(Albrecht 2015: 76, Fig. 5)



Functional categories by gender (per one million words)
(Albrecht 2015: 75, Fig. 4)



Usage of epistemic adverbs in BA and MA theses per 10,000 words
(Beyer 2015: 93, Fig. 3)



Sentence subject in *may*-clauses per 100,000 words (Küchler 2015: 109)

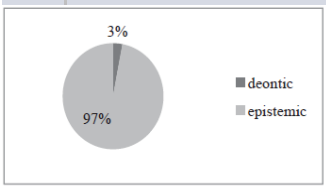


Figure 2: Frequency of *may* in the *ChemCorpus* TP (N=399)

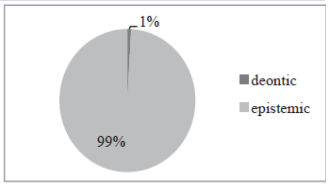
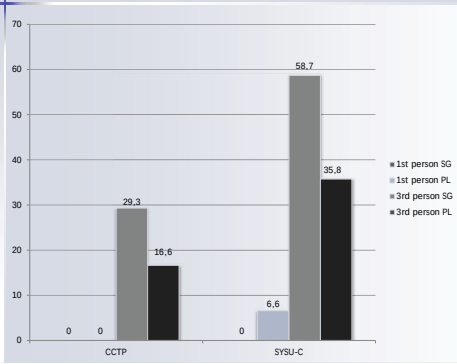


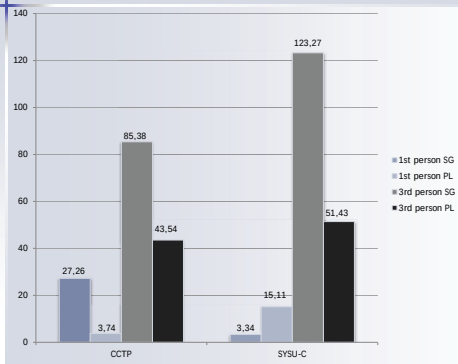
Figure 3: Frequency of *may* in the *SYSU-C*orpus (N=901)

- This may be one reason for the implementation of the New Economic Policy (NEP) in 1971. (CTP08FBAC_128)
- Graduation is concerned with the resources by which the force or intensity of an utterance may be raised or lowered. (CC11FMAT_3)
- They may lack time or knowledge ... (CC11FMATP_83)
- Maybe, it is possible to think that both systems may appear to be irrelevant to each other. (CTP12FBALJR_23)

Sentence subject in *may*-clauses per 100,000 words
(Küchler 2015: 110, Fig. 5)



Sentence subject in *will*-clauses per 100,000 words
(Küchler 2015: 114, Fig. 10)



Meanings of *will* per 100,000 words
(Küchler 2015: 113, Fig. 9)