Corpus Linguistics
DGfS & GLOW Summer School
Micro- and Macrovariation
Stuttgart, 2006
Anke Lüdeling
anke.luedeling@rz.hu-berlin.de

organisation
- course plan (→ handout)
- all materials will be available online at
  http://www2.hu-berlin.de/korpling/lehre/SummerSchoolStutt.php/
- mini-projects (in groups):
  ten minute presentation plus short written summary
- you can contact me via email any time

Anke Lüdeling, DGfS & GLOW Summer School,
Stuttgart, Aug 2006

corpus linguistics
- is concerned with
  - design
  - processing (architecture and annotation)
  - evaluation
- of corpus data, where a corpus is
  “A collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language.“
  http://www.ilc.cnr.it/EAGLES96/corpintr/corpintr.html

Anke Lüdeling, DGfS & GLOW Summer School,
Stuttgart, Aug 2006

today
- different kinds of data for different
  linguistic research questions
- brief history of corpus linguistics

Anke Lüdeling, DGfS & GLOW Summer School,
Stuttgart, Aug 2006

linguistic data
- where do linguists get the data they need
to test hypotheses/theories?
  - introspection
  - psycholinguistic experiments
  - neurolinguistic experiments
  - questionnaires
  - corpora

Anke Lüdeling, DGfS & GLOW Summer School,
Stuttgart, Aug 2006

linguistic data
- the research question and the theoretical
  framework determine the kind of data hat can be
  used (the mantra ;-))
  - in many cases different kinds of data need to be
  integrated
- methodology/data issues are discussed a lot in
  recent years, see http://www.sfb441.uni-tuebingen.de/index-engl.html, several
  conferences like Linguistic Evidence, QITL, etc.,
  Bod, Hay & Jannedy 2003, Kepser & Reis 2005, ...

Anke Lüdeling, DGfS & GLOW Summer School,
Stuttgart, Aug 2006
introspection/intuition
- arm-chair linguistics ;-)  
- generative tradition, competence model  
- research question: which complex expressions (phrases, sentences, words, ...) can be produced by the internal grammar of a native speaker of a language?  
- grammaticality judgments, yes/no  

psycholinguistic experiments
- research questions: how are linguistic data stored and accessed?  
- storage and processing models, interaction with other cognitive tasks  
- elicitation tasks, reaction time experiments, eye tracking, errors, ...

neurolinguistic experiments
- research questions: where are linguistic data stored and accessed?  
- storage and processing models, interaction with other cognitive tasks  
- EEG, ERP, imaging techniques, ...

Lynn Frazier's class  

Bomkessel & Schlesewski's class

questionnaires
- different research questions  
  - structuralist/generative fieldwork to find grammatical forms  
  - judgment tasks (yes/no, magnitude estimation, ...)  
  - ...

- an experimental technique (issues of representativity, filler items etc.), often used to verify specific hypotheses, often used for quantitative studies

Bomkessel & Schlesewski's class

corpora
- research questions: see following slides  
- collections of texts  
  - nowadays mostly electronic, but see history  
  - texts usually produced for independent purposes (authenticity) – collection (design) and evaluation methods dependent on the research question

corpora – research questions
- many areas traditionally use corpora  
  - historical linguistics  
  - sociolinguistics  
  - dialectology  
  - lexicography  
  - language acquisition research  
  - (computational linguistics/natural language processing)  
- other areas have only recently started using corpora  
  - generative/theoretical linguistics
historical linguistics

- (there is no other kind of data ;-) 
- question: what was a variety (at least that portion of the language that survived) like at a specific point in time – qualitative and quantitative study of a 'synchronic' corpus (many papers ...)
- question: how can language change be described/modelled – comparison of corpus data from different points in time (also short-term/recent change!)
- Eythorsson’s class

historical linguistics – example

- automatic calculation of similarity trees to find out about language relationships between historical varieties of German (Lüdeling 2006)

small case study: Lord’s prayer

- fader ist usa firio barno, thu bist an them hohen himilo rkie. Aitsachsich (9. Jhd., Helian)
- fater unser, thu thar bist in himile, Ahd. (9. Jhd., Tatian)
- Got vater unser, dâ du bist In dem himelriche gewaltic alles des dir ist, Mhd. (ca. 1200, Reinmar van Zweter)
- Vnser vater ynn dem hymel. Fnhd. (1422, Luther)
- Vater unser, der du bist im Himmel, Nhd.

‘intuitive’ distance

<table>
<thead>
<tr>
<th></th>
<th>AS</th>
<th>AHD</th>
<th>MHD</th>
<th>FNHD</th>
<th>NHD</th>
</tr>
</thead>
<tbody>
<tr>
<td>AS</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>AHD</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>MHD</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>FNHD</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>NHD</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

experiments

- word lists (Hochmuth 2004)
- syntactic similarity using the TIGER format (Brants et al. 2002)
- visualisation of distances using PHYLIP phylogeny software

word lists

- translated to SAMPA
- manually aligned
- grapheme correspondences, SAMPA correspondences
- several string comparison methods (edit distances)
- phylogenetic methods for clustering
hierarchical clustering, feature-weighted Levenshtein distance

syntactic distance
- transformation of TIGER graphs to trees, crossing edges ignored (order of terminal nodes)
- calculation of distances between corresponding sentences using treediff (Shasha, Wang & Zhang)
- sentence distances combined to text distance

phylogenetic trees

<table>
<thead>
<tr>
<th></th>
<th>AS</th>
<th>AHD</th>
<th>MHD</th>
<th>FNHD</th>
<th>NHD</th>
</tr>
</thead>
<tbody>
<tr>
<td>AS</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AHD</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MHD</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>FNHD</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>NHD</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

'intuitive' distances vs. Lord's prayer

Lord's Prayer
problems
- too little data
- the Lord's Prayer is a translation and formulaic
  - fater unser, thu thar bist in himile, Ahd.
  - Vater unser, der du bist im Himmel, Nhd. vs.
  - Unser Vater im Himmel

sociolinguistics
- question: what characteristics does a given sociolect/register/… have (wrt pronunciation, lexis, syntax, …)?
- question: how do sociolects/registers/… differ?

historical linguistics/ sociolinguistics/ descriptive grammars
- we want:
  - qualitative and quantitative description of a language wrt all linguistic levels
  - perhaps extrapolation of results to a larger body of language (OHG, London Teenager Talk, …)
- we need:
  - reliable corpora with good documentation, reliable annotation with good documentation
  - search and evaluation techniques
  - mathematical models

corpus-based methods: limits & chances
- corpora (comparability, availability)
- annotation
  (linguistic levels, tag sets, guidelines, tools)
- similarity measures
- models (trees, nets, ?)
- base the calculation of language relationships on more than just phonology and morphology
- frequencies can be used

descriptive grammars
- question/task: description and classification of basic elements and combinations in a language (and perhaps determine frequency information)

psycholinguistics
- question/task: find frequency information for words/morphemes/readings etc. in order to design experiments
  (reaction times are correlated with frequency)
- question/task: find typical contexts for a given word
- we want: large general corpora or corpus-based lexicons with frequency information, good annotation
**lexicography**

- question/task: find readings/typical contexts/frequencies of a word, find good examples
- question/task: find collocates to a given word
- we want: large general corpora or special corpora (for special lexicons), good annotation and evaluation methods, mathematical models for collocation detection
- WordSketch (Kilgarriff), Evert 2005

**linguistic data**

<table>
<thead>
<tr>
<th>Introspection</th>
<th>Experimental data</th>
<th>Corpus data</th>
</tr>
</thead>
<tbody>
<tr>
<td>competence: what is grammatical?</td>
<td>how is language processed?</td>
<td>what occurs?</td>
</tr>
<tr>
<td>(formal) production system that produces all (and only the) grammatical expressions of a language</td>
<td>model that describes storage of linguistic units and the way those are addressed and processed, neurological models</td>
<td>model that describes those linguistic units and combinations and their distribution in a corpus</td>
</tr>
</tbody>
</table>

**history of corpus linguistics**

- text collections are used already in the 19th century (and earlier) to
  - describe language change
  - illustrate statements about grammar
  - document language acquisition
  - compile dictionaries
  - compare languages

**historical examples**


**structurelism**

- synchrony
- spoken language
- American structuralism:
  - Boas, Sapir, Bloomfield, Harris
  - influential at least until the 1960s, terminology and empirical methods still in use
  - corpus-based – corpora often small, systematically collected through elicitation (almost questionnaire studies), no quantitative studies
  - non-European languages (native American languages)
structuralism

- debate: should a grammar describe the collected fragment (the corpus) or is it possible to extrapolate from that to the whole language?
- basic idea: a language is finite – if you collect enough data you could in principle collect everything

generative theory (Chomsky)

- new research goal: people understand and produce infinitely many sentences/complex expressions. It is therefore not interesting to describe a finite sample – the real goal is description of the underlying production system
- competence vs. performance
- i-language vs. e-language

generative theory

- a corpus is a collection of performance data which are influenced by all kinds of non-linguistic factors
  - since it is not possible to abstract away from these extra-linguistic factors a corpus cannot be used to find a competence model
  - introspection is the only way one can use to distinguish between grammatical and ungrammatical utterances

grammaticality and corpus data

<table>
<thead>
<tr>
<th>Grammatical</th>
<th>Ungrammatical</th>
</tr>
</thead>
<tbody>
<tr>
<td>occurs in a corpus</td>
<td>immer, Wirtschaftskrise, nach 14 Jahren Kohl</td>
</tr>
<tr>
<td></td>
<td>immer, letztendlich, unkaputtbar, ich habe fertig, ...</td>
</tr>
<tr>
<td>does not occur in a corpus</td>
<td>NPs with 27 genitive attributes: das Haus der Großmutter der Schwester des Verwalters der ...</td>
</tr>
<tr>
<td></td>
<td>&quot;Some sentences won't occur because they are obvious, others because they are false, still others because they are impolite.&quot; (Chomsky 1962, 159)</td>
</tr>
</tbody>
</table>

Chomsky on corpus linguistics

"It doesn't exist"

(Chomsky in an interview, answering a question by Bas Aarts "What do you think of corpus linguistics?", 2001)
corpus linguistics after the 1950s

- some linguistic areas always used corpora and continued to do so (dialectology, historical linguistics, …)
- computational linguistics and psycholinguistics have interest in machine-readable corpora because they need frequency data
- theoretical linguistics: corpus-based work marginalized
- almost no discussion of empirical questions/standards

early machine-readable corpora

- Roberto Busa: corpus of medieval philosophy texts (project with IBM 1949 – 1967), concordancer (this later lead to: Thomae Aquinatis Opera Omnia cum hypertextibus in CD-ROM)
- other work on historical texts (Greek Bible and other texts)
- Morton’s authorship detection
- Juillands ‘mechanolinguistics’

early machine-readable corpora

Memories of the early days are all of paper tape. It waved in and out of every machine, it dried and then cracked and split or it got damp when it lay limp and then sagged and stretched. Sometimes it curled round you like a hungry anaconda, at others it lay flat and lifelike and would not wind. Above all it extended to infinity in all directions. A Greek New Testament, half a million characters, ran to a mile of paper tape, and the complete concordance of it ran to seven miles (Morton 1980, 197).
(zitiert nach http://info.ox.ac.uk/ctitext/history/pioneer.html)

second generation corpora

- much larger (> 100 m tokens)
- standardization initiatives
- networks
- much more corpus-based research
- the gap between corpuslinguists and theoretical linguists is closing
"corpus linguistics is not a branch of linguistics, but the route into linguistics".

Michael Hoey, remark at TALC 1998