Negative Polarity Items
Corpus Linguistics, Semantics, and Psycholinguistics
Day 2: Corpus Linguistics: Qualitative Research

Frank Richter¹, Janina Radó¹, Manfred Sailer²

¹Universität Tübingen
²Universität Göttingen

ESSLLI 2008, Hamburg
Summary of possible answers

- different theories have different focus, look at different classes of data.
- difficulties of comparing theories
- limitations of introspective data (variation, context-dependent licensing, comparison with other phenomena)
- status of the data is not clear at all
Why empirical research?

- Licensor:
  Can we find more contexts? Are all contexts equally fine?

- Licensee:
  Can we find more NPIs? What classes of NPIs are there?

- Relation:
  Can we test whether the relation between the NPI and the licenser patters with other linguistic relations?

- Status:
  Can we test the status of sentences with unlicensed NPIs?
Empirical methods

- corpus linguistics: usage data
  - more data on the usage of known NPIs
  - NPI classification based on usage possible?
  - usage data essential for context-dependent readings
  - find new NPIs

- psycholinguistics: judgment and processing
  - NPI classification
  - investigation of intervention
  - answer to the status question
Summary of day 1

What we saw yesterday:
- Four questions on NPIs
- Four attempts to answer them

Conclusion
- diverse theoretical approaches with different predictions
- empirical basis still not settled.

Outlook
- Tuesday & Wednesday: corpus linguistics
- Thursday & Friday: psycholinguistics
Assignment Day 1

- **Aim:** Find six different NPIs in your native language that belong to at least three different syntactic categories.

- **Method:**
  1. Pick six to eight NPIs from the file: `english-npi.pdf`
  2. Translate the items into your native language.
  3. Test whether the translations are NPIs as well.

  Diagnostic environments:
  (i) Can the item occur in a clause whose subject is "nobody"?
  (ii) Is the sentence grammatical when you use the subject "Pat" instead?

  example:
  (i) **Nobody** had the ghost of a chance of getting the job.
  (ii) * Pat had the ghost of a chance of getting the job.

- **Mail your results by tomorrow 11am to**
  manfred.sailer@phil.uni-goettingen.de
Your NPIs

Languages: Dutch, English, German, Hungarian, Italian, Irish

Problems:

- idioms don’t translate: *x is as black as x is painted*
- item is no NPI in the translation
- the subject is part of the NPI: *not a doggoned thing has happened.*
- difficulties to tell apart idiomatic from non-idiomatic reading (*have an idea*)
- German: particles *überhaupt, fei,*
- items that occur in question: *zur Hölle*
- negation must be in the same constituent: *(k)ein Vergleich*
Corpus

Definition
A corpus is a collection of written or spoken utterances in one or more languages. The data in a corpus are digitalized. The corpus consists of primary data (texts or sequences of utterances) together with meta-data describing the texts and linguistic annotation of the data. taken from Lemnitzer and Zinsmeister (2006), p. 7
Advantages of and objections against copora

- **Advantages**
  - data in context
  - natural data
  - quantitative statements possible
  - often uncovers new types of data
  - direct access to native speakers less necessary

- **Objections**
  - no negative evidence possible
  - ungrammatical data in the corpus
  - not all relevant constructions present
  - most data are of a similar kind

- **To avoid problems:**
  - representative corpus
  - metadata
  - linguistic annotation (?)
Corpora in linguistics (Lemnitzer and Zinsmeister, 2006)

- corpus-based, qualitative research: is a certain combination of words possible in the language?
- corpus-based, qualitative-quantitative research among two variants of a word, which one is the preferred?
- corpus-driven research: extract all words that occur in a given environment
Corpus-based, qualitative research

- explore word order phenomena
- Generative Grammar
- aim: find sentences that can be subjected to introspective judgments
- input: linguistically annotated corpus, collection of relevant example sentences
- output: individual example sentences
- no statistics!
- interpretation of the findings from the theory
- applied in lexicography, theoretical linguistics
Corpus-based, qualitative-quantitative research

- collocation analysis
- British contextualism (Firth, Sinclair)
- aim: explore the usage of an item
- input: bare text corpus
- output: collocator-collocant pairs
- frequency, but no statistics!
- linguistic interpretation of the attested examples
- applied in lexicography, foreign language teaching, translation science
Corpus-driven, quantitative research

- n-gram analysis, latent semantic analysis
- quantitative language processing
- aim: extract patterns from the corpus using little to no linguistic knowledge
- input: bare text corpus
- output: n-grams with frequencies
- statistical model is very important
- applied in information retrieval, speech processing
Corpora

- www.essex.ac.uk/linguistics/clmt/w3c/corpus_ling/content/

- German:
  - Digitales Wörterbuch der deutschen Sprache (DWDS):
    www.dwds.de
  - Publically available corpora of the Institut für Deutsche Sprache, Mannheim:
    http://www.ids-mannheim.de/cosmas2/web-app/

- English:
  - British National Corpus (BNC): info.ox.ac.uk/bnc
  - Cobuild data base:
    www.collins.co.uk/Corpus/CorpusSearch.aspx
The web as a corpus

- useful for qualitative studies
- huge! With context!
- problems: reliability of the data? Long-term availability of the data?
- no quantitative statements possible
- indicate:
  - search engine
  - search pattern
  - date of the query
  - URL
  - example
Corpus-based, qualitative research on NPIs

- typical questions: Can an NPI occur in a particular construction?
- collect and classify occurrences of an NPI
Can an NPI occur in a particular construction?

- intervention effect:
  
  *I don’t think Pat/ every student skipped any ESSLLI reception.*

- google, August 4, 2008, query: “I don’t think everyone ever”

- Not a relevant hit:
  
  *I’ll admit that I don’t think every single song they’ve ever written is perfect, but when they write a gem, its a good one. Their latest release "Living*"

  www.punkrockparents.com/sensefieldlivingoutsiderev.htm

- Relevant hit (native?):
  
  *So I don’t think every single Humvee will ever be supplanted. How many are replaced are decisions that will be made by the operations commanders,*

Collect and classify occurrences of an NPI (Sinclair, 2004)

- Cobuild corpus  
  (www.collins.co.uk/Corpus/CorpusSearch.aspx)
- example: *budge*
Results of the qualitative analysis of *budge*

- irrelevant data:
  - typo (*budget*)
  - proper name

- typical NPI data:
  - *did not*, *would not*, …
  - *refuse to*, *refusal to*
  - *they weren’t prepared to*
  - negative inversion: *not another step will I b.*
  - *neither side seems particularly ready to b.*

- challenging data:
  - no clear licenser:
    *their vain attempts to budge even one of the great monoliths*
  - context to small:
    *because this is just an excuse to toss in a budge factor.*
Evaluation

- interesting data
- typical usage patterns
- What about other typical contexts?
  Explore occurrence gaps by construction-specific searches in other corpora! (and by introspection)
- Result: qualitative profile of the NPI.
Collection of Distributionally Idiosyncratic Items


