From recording the past to predicting the future?
On the role and relevance of linguistic abstraction for corpus-based analysis

Detmar Meurers
University of Tübingen

Herrenhausen Conference
“(Digital) Humanities Revisited – Challenges and Opportunities in the Digital Age”
Hannover, December 5-7, 2013
Guiding question of this section: Digital Humanities — What kind of knowledge can we expect?

Linguistics studies
- how language is acquired by individuals
- how languages change over time and influence each other
- how form and meaning interact in language as a system
- how language use correlates with personal identity, . . .

The digital world provides increasingly large sets of data:
- corpora collected in different contexts (news, subtitles, . . .)
- learner corpora (e.g., 76k learners in EFCamDat)
- historical corpora
Introduction
New data sources driving research

▶ The increasing size and representativeness of digital language data supports insights into human language.
  ▶ Frequencies based on TV subtitles are best predictor of human word processing abilities (Brysbaert et al. 2011a,b).
    ▶ Representativeness matters, not size as such (size above 20–30 million words of little value, Brysbaert & New 2009).

▶ At the same time, with the availability of large corpora, language often seems to be reduced to surface forms.

▶ Language as a bag of words is also popular in tools:
  ▶ Latent Semantic Analysis used for real-life essay grading
  ▶ Statistical Machine Translation based on bilingual corpora
Introduction

Steinbeck’s cannery row, or: counting surface forms is fishy

▶ Relying on surface forms misses relevant underlying classes.
▶ But corpora can be annotated with classes, can’t they?
Annotating corpora

- Where do linguistic categories come from?
- Categories result from generalizations, which establish labels for sets of observable properties.
  - linguistic categories rooted in analysis of Latin, Greek
  - recent categories (e.g., sentiment analysis) established using annotation schemes and reference corpora
- Example: Three sources of evidence for parts-of-speech
  1. *I was surprised by the word of the day.*
     - lemma: *of* ⇒ preposition
  2. *There is a lot of construction going on.*
     - morphology: *-ion* ⇒ noun
  3. *The old man left.*
     - distribution: adj __ verb ⇒ noun
Categories appropriate for learner language?

(Díaz Negrillo, Meurers, Valera & Wunsch 2010)

(4) RED helped him during he was in the prison.
   ▶ lemma: preposition
   ▶ distribution: conjunction

(5) one of the favourite places to visit for many foreigners.
   ▶ lemma: adjective
   ▶ distribution, morphology: noun

(6) to be chosen for a job
   ▶ lemma: noun or adjective
   ▶ distribution, morphology: verb

▶ A single POS tag from a standard native tagset fails to systematically identify properties of learner language.
▶ “Robust” categorization can hide relevant characteristics.
On the nature of categories

- **Comparative fallacy**: “mistake of studying the systematic character of one language by comparing it to another” (Bley-Vroman 1983, p. 6)

- Issue as such is quite general:
  - Eurocentrism in field work (Gil 2001)
  - hermeneutic circle: interpretation of text in context

⇒ To provide access to the abstractions relevant for a range of research questions, one needs
  - multiple types of annotation,
  - supporting different levels of granularity,
  - and robust category assignment should be based on explicit target hypotheses (Lüdeling 2008).
Explicit operationalization as an opportunity

- How can these annotation layers be obtained?
  - automatic tools (taggers, parsers, classifiers)
  - crowd sourcing linguistic annotation:
    - requires rethinking linguistic expert knowledge as empirical tests which can be carried out by anyone
    - cf. new methods in linguistic field work (Tonhauser 2012)

- Digital Humanities can be viewed as an opportunity
  - to revisit the underlying concepts and categories
  - revise and fully operationalize them, and
  - highlight their empirical value and explanatory potential.
An experimental testbed for linguistic abstraction

- How can we find out more about the informativeness of surface forms and linguistic abstractions?
  → Set up a classification experiment which allows us to quantify the impact of different features.
    - supervised machine learning:
      - study record of the past: train on labeled data
      - test model predictions of “future”: classify unseen data

- Test case: Identify native language given non-native text.
  - *Transfer is the influence resulting from similarities and differences between the target language and any other language that has been previously acquired.* (Odlin 1989)
  - involves all levels of language (lexis, grammar, . . .)
  - core topic of Second Language Acquisition research
Two strands of experiments

- Data-driven approach (with Serhiy Bykh):
  - from surface forms to part-of-speech

- Theory-driven approach (with Julia Krivanek):
  - from syntactic alternations to data-informed patterns
Data-driven approach

Setup

- International Corpus of Learner English (Granger et al. 2009)
  - argumentative essays written by higher intermediate to advanced learners of English
  - subcorpus with seven native languages: Bulgarian, Czech, French, Russian, Spanish, Chinese, Japanese
  - 95 texts per language, between 500 and 1000 words long
  - extract all sequences of words occurring at least twice
    - 67,905 n-grams of length 2–28
  - use each such recurring n-gram as a binary feature:
    - 1 if it occurs in the text, 0 if not
  - trained a classifier (SVM) on 70 texts for each language
Data-driven approach

Surface-form results

- Result on held-out test set (25 texts per language):
  - classification accuracy: 87.4%
  - random baseline (7 languages): 14.3%
  - Wong & Dras (2009): 73.7%

- What happens if we abstract away from the word features
  - to words with the same part-of-speech?
  - to any words occurring within recurring frame?
Data-driven approach
Example for feature abstraction

- Part-of-speech abstraction:
  - 3-grams:
    - each JJ it
    - environment IN which
    - family RB at
    - few NNS later
  - 4-grams:
    - they VBP IN the
    - for JJ NN to
    - different NNS IN view
    - would VB RB longer

- Non-linguistic abstraction:
  - 3-grams:
    - each * it
    - environment * which
    - family * at
    - few * later
  - 4-grams:
    - they * * the
    - for * * to
    - different * * view
    - would * * longer
Results

- Generalization to linguistics classes improves the results, whereas non-linguistic abstraction does not.

- Success, but hard to interpret features in terms of transfer!
Observing choices in the linguistic system

- Word-based surface features encode form and meaning.
  - This requires very high number of features to be applicable to unseen data, across domains/topics.

- Can we abstract away from the meaning to be expressed to choices in the linguistic system?
  - Study where the linguistic system provides multiple ways to express the same meaning. (cf. variationist sociolinguist.)

- How about valence alternations (Levin 1993)?
  1. *He gave the book to John.* “Dative Alternation”
  2. *He gave John the book.*
Theory-driven approach

- Task: binary classification into non-native vs. native
- Corpus used: 720 documents evenly drawn from
  - Chinese English from ICLE (Granger et al. 2009)
  - native English from LOCNESS corpus
- Features:
  - 21 alternation which can reliably be identified automatically given syntactic annotation (a fifth of Levin’s alternations)
  - encode document as relative frequency of choices made
Qualitative analysis
Locative Preposition Drop Alternation is distinctive

L1 Chinese

- **V-PPloc**: 0.15
- **V-NP**: 0.85

L1 English

- **V-PPloc**: 0.37
- **V-NP**: 0.63

(Martha climbed up the mountain.)

(Martha climbed the mountain.)
Qualitative analysis

Dative Alternation is indistinctive

L1 Chinese

L1 English

V-NP-NP  V-NP-to-NP

(He gave John the book.)
(He gave the book to John.)
Theory-driven approach
Results . . . and improvements using a data-driven twist

- Result: 63.33% classification accuracy
  - Alternations good in theory, but don’t occur often enough!

- Can we infuse more data-driven life into the alternations?
  - for each verb, record its selection patterns in the corpus
  - define classes consisting of all verbs with same patterns
  - significantly improves results: 72.5% accuracy

- Combination of theory & data-driven perspective is viable
  - applicable to morphological choices (Krivanek & Meurers 2013)
  - next steps:
    - systematically explore range of choices in linguistic system
    - interpret findings in terms of a theory of Transfer
Summary

- Large scale digital data
  - provides opportunities for analyzing language,
  - but also a clear danger of only analyzing the surface.

- There is a need to preserve
  - genuine research questions rooted in the field
  - interpretation of data informed by classes and context

- To support a range of research questions, corpora need
  - multiple annotation layers, for which
  - automatic annotation and crowd sourcing requires
  - revisiting and operationalizing the categories and interpretations underlying the field of study.

- Experimental test beds can be set up
  - to quantitatively validate conceptual advances
  - in a way that supports qualitative analysis of features.
Outlook

- Complementing the Digital Humanities (pre)occupation with surface-near exploration of large-scale data,
  - it increasingly offers the opportunity to enrich the data
    - with the classes, structure, and context needed
    - to address (further) research questions in the humanities.
References


Introduction

Counting words without context is no help

- Negative polarity items such as *any* typically occur in the context of negation, but they do not express the negation.
- Counting words without context leads to misinterpretation.