Linguistic Modeling and Analysis

Anna Shadrova, Martin Klotz, Anke Lüdeling

Humboldt-Universität zu Berlin
Institut für deutsche Sprache und Linguistik

DGfS 2021

Funded by
Deutsche Forschungsgemeinschaft
German Research Foundation

Linguistic modeling and analysis

In corpus-linguistic research

- we aim to model and understand higher-level linguistic concepts
- we want to explain distributions of linguistic features in naturalistic or semi-naturalistic data, meaning we want to quantify
- the analysis itself is the result (unlike in applied computational linguistics or information retrieval, where external usability can provide proof of relevance)

Linguistic modeling and analysis

In corpus-linguistic research

- we aim to model and understand higher-level linguistic concepts
- we want to explain distributions of linguistic features in naturalistic or semi-naturalistic data, meaning we want to quantify
- the analysis itself is the result (unlike in applied computational linguistics or information retrieval, where external usability can provide proof of relevance)
  - high demands on the model of analysis (must be able to capture fine-grained and fuzzy edged categories, ambiguity)
  - high demands on accuracy (any divergence between analyses can imply relevant theoretical differences)
  - at least moderate demands on data size (must be large enough for quantitative analysis)

Varying degrees of abstraction and ambiguity

Linguistic features can be expressed as

- surface-near and largely unambiguous to a human reader – easily automatable and quite accurate (lemmatization, morphological tagging)
- surface-near and somewhat ambiguous – less easily automatable and less accurate (POS-tagging, phonetic analysis)
- surface-near and structurally ambiguous – less easily automatable and even less accurate (syntactic parsing)
- abstract, i.e. surface-ambiguous, but largely unambiguous to a human reader (semantic categorization, named entity recognition, anaphora resolution)
- abstract, i.e. surface-ambiguous and highly ambiguous even to a human reader (rhetorical structures, argument mining)

Resources of present-day corpus linguistics

- Access to large amounts of data:
  - several large corpora with or without more surface-near annotations
  - many smaller, well-controlled, and deeply annotated corpora developed w. r. t. a specific research question
  - masses of reusable digitized (text) data from other contexts

Resources of present-day corpus linguistics

- Access to large amounts of data:
  - several large corpora with or without more surface-near annotations
  - many smaller, well-controlled, and deeply annotated corpora developed w. r. t. a specific research question
  - masses of reusable digitized (text) data from other contexts
- Giant leaps in computational power even on simple devices such as a laptop:
  - (simple) deep learning
  - computation involving complex graphs
  - multifactorial analyses
  - computation of complex distributions in general
  - ...

Resources of present-day corpus linguistics

- Enormous developments in computational linguistics, especially in application-oriented approaches based on machine learning:
    - collaborative efforts: shared tasks, openly accessible models and libraries
    - large-scale language models and word embeddings (Brown et al., 2020; Devlin et al., 2019)
    - knowledge graphs and semantic databases (WordNet, Wikidata, Google Knowledge Graph, BabelNet, ...)

Modeling and analysis: computational linguistics

Computational linguistics has recently made the most prominent advances in its application/task-oriented branches:

- models are developed functionally and pragmatically with major concessions to the underlying linguistics:
    - language or architectures are modeled towards a task (rather than a question) → models are often context-sensitive and task-specific
    - quantitative accuracy > analytical depth
        - the underlying machine learning models are often very hard to interpret (Harbecke, 2021)
        - performance is increasingly gained from surface forms only, analysis is of little concern

Modeling and analysis: computational linguistics

Computational linguistics has recently made the most prominent advances in its application/task-oriented branches:

- models are developed functionally and pragmatically with major concessions to the underlying linguistics:
    - language or architectures are modeled towards a task (rather than a question) → models are often context-sensitive and task-specific
    - quantitative accuracy > analytical depth
        - the underlying machine learning models are often very hard to interpret (Harbecke, 2021)
        - performance is increasingly gained from surface forms only, analysis is of little concern

→ more and more diverse but shallow data rather than a deeper understanding of language

Theses for discussion

Thesis 1: Due to its increasing focus on surface forms, NLP does not suffice or is not employed in helpful ways for most linguistic research questions at present.

How do we model beyond the surface?

1. We might face different problems:
    - Theory is well-developed, annotation is application – but not easily trainable
    - Theory is not yet well understood, not yet automatable

How do we model beyond the surface?

1. We might face different problems:
    - Theory is well-developed, annotation is application – but not easily trainable
    - Theory is not yet well understood, not yet automatable

→ But how do we decide which one is the case...

- in rhetorical structure theory (Mann and Thompson, 1987)?
- in information status/structure (Riester and Baumann, 2017)?
- in anaphora resolution?
- ...

Challenges to linguistic modeling

- Structural ambiguity: the same data looks different within a single model depending on the context or interpretation
- Model ambiguity: the same data looks different through different models (can become an issue if analysis aims at generalizability or reusability)
- Incomplete, insufficient, or ambiguous modeling: the models we use may not always provide uniquely definable mappings, categorizations, or analyses for all cases we encounter

Linguistic theory and annotation | example

Research question: How do learners of German as a Foreign Language (and native speakers of German) use complex nouns? Do they use word-formation rules productively?

This research is done on the Falko corpus (hu-berlin.de/falko) in the context of Project C04 of SFB 1412 Register, see Lukassek et al. (2021).

Linguistic theory and annotation | example

- morphological descriptions and reference books: complex nouns can be the result of compounding, derivation, and (sometimes) conversion; everything else is described as rare and mostly irregular
- the researcher develops guidelines

Linguistic theory and annotation | example

- morphological descriptions and reference books: complex nouns can be the result of compounding, derivation, and (sometimes) conversion; everything else is described as rare and mostly irregular
- the researcher develops guidelines
- hm. what should be done with
  - transparent and less transparent non-native words (*Kriminalität* "crime", *Knowhow*, *Abitur* "high school diploma")?
  - 'irregular' noun formation processes (*Rede* "speech", *Satz* "sentence")?
  - syntactic transpositions (*das Leben* "life", *der Verletzte* "the injured")
  - different types of 'irregular' or (synchronically) partly obscure words such as pluralia tantum (*Eltern* "parents", *Kosten* "costs") or synchronically opaque formations (*Geschlecht* "gender")
  - structurally ambiguous words (*Stellungnahme* "statement")?

Linguistic theory and annotation | example

It becomes clear that the answer to these questions (and the subsequent annotation) determines which research questions can be addressed and which questions cannot be addressed

→ the annotation process and the repeated discussions lead to a much clearer understanding of the phenomenon – always in the light of the research question

Linguistic theory and annotation | typical annotation procedure

Starting from a research question

- the researcher wants to classify the data according to (some version of) of given linguistic model
- she develops guidelines
- the data is not well-behaved, it resists easy classification; it demands a better analysis
  - often (sadly): the researcher makes the data fit
  - sometimes: the researcher performs a detailed analysis the annotation process leads to a deeper understanding of the phenomenon and, subsequently, to better guidelines

Theses for discussion

Thesis 2: Often, our linguistic concepts are not well-defined and easy to operationalize – iterative annotation, especially of higher level concepts, *is* linguistic modeling.

Similarly, approximating phenomena additively through the combination of various surface features and measures, is also an implicit specification – i.e. modeling – of the higher-level linguistic concept that we try to approach.

Quantification

Corpus linguistics strives to quantify and to automate

- for exactness and reliability of the analysis
- for gradual comparison between groups and factors
- in order to capture dynamics

Complexity of linguistic variability

- High degree of variability by many factors:
  - (somewhat) stable groupwise inter-individual: age; geographic, social, and linguistic background, among others
  - (somewhat) stable random inter-individual (stylistic preferences, individual habits of expression)
  - (somewhat) stable intra-individual: register, modality
  - dynamic intra-individual: cognitive and psychological state, e.g. motivation, fatigue
  - dynamic path dependency: dialogue dynamics such as priming and alignment
  - (somewhat) stable path dependency: entrenchment, acquisition/attrition, language change

Example: RUEG corpus (Wiese et al., 2020)

task-based heritage and majority language corpus

- bilingual speakers in majority and heritage language
- monolingual speakers as controls for comparison with heritage speakers
- two levels of formality
- two modalities
- adolescents and adults
- male* and female*

Example: RUEG corpus (Wiese et al., 2020)

task-based heritage and majority language corpus

- bilingual speakers in majority and heritage language
- monolingual speakers as controls for comparison with heritage speakers
- two levels of formality
- two modalities
- adolescents and adults
- male* and female*

→ Differences in linguistic realizations are found across factors: $2^6 = 64$ subgroups

Example: RUEG corpus (Wiese et al., 2020)

task-based heritage and majority language corpus

- bilingual speakers in majority and heritage language
- monolingual speakers as controls for comparison with heritage speakers
- two levels of formality
- two modalities
- adolescents and adults
- male* and female*

$\rightarrow$ Differences in linguistic realizations are found across factors: $2^6 = 64$ subgroups ... per language!

Theses for discussion

Thesis 3: To yield accurate results, quantitative corpus analyses must account for the high degree of complexity, interaction, and variability in linguistic data.

We need to collect sufficient amounts of parallel data, and, since time constraints apply, quantitative frameworks that do not rely on very large data.

Linguistic modeling and analysis

- Step 1: Identification and accounting for all relevant influential factors in data collection
- Step 2: Full, linguistically valid, and reliable analysis of all cases in the dataset (iterative annotation = linguistic modeling)
- Step 3: Quantitative modeling: Mapping of linguistic to quantitative model
  → But what and how do we quantify?

(In practice, these steps are not usually neatly separable)

Quantification of morphological productivity

- There are many measures for different aspects of morphological productivity. Most are based on type-token distributions in a corpus (Baayen, 2001; Zeldes, 2012).

  But do those measures measure what we want to measure?

Quantification of morphological productivity

- There are many measures for different aspects of morphological productivity. Most are based on type-token distributions in a corpus (Baayen, 2001; Zeldes, 2012).
  But do those measures measure what we want to measure?
- We might also need models that refer to degrees of transparency and lexicalization
  → properties of the speaker and not of the corpus (corpus counts can only approximate this)

→ This information would require types of models that are able to map probabilities.

Formalization and quantification of morphological productivity

...but regarding the research question above, it might be more relevant to look at (regularly-formed and irregularly-formed) morphological families:

- *setzen - sitzen - Setzung - Satz - einsetzen - Einsatz - ...*
- *kriminell - kriminalistisch - Kriminalität - Autorität - Anonymität - ...*

Modeling such families would require a graph model, which is an entirely different formal model equipped with its own types of quantification

Quantitative comparison of rhetorical structures in use (Wan, in prep.)

- If we want to quantitatively compare the use of rhetorical structures in L1 and L2 writing, we need a measure of similarity for rhetorical structures

→ So how do we quantify tree similarity? For example: Is the same embedding level with a different label more or less similar to the same label at different embedding level?

Quantitative comparison of degrees of coselectional constraint (Shadrova, 2020)

- If we want to compare the “nativelikeness” of learners at different stages of acquisition, we need an operationalizeable and quantifiable concept of “nativelikeness”, for example: degree of coselectional constraint on verb-argument structures

Quantitative comparison of degrees of coselectional constraint (Shadrova, 2020)

But what is coselectional constraint?

- is it: the strength of lexical association between words? I.e. “the more strongly individual words are associated, the more idiomatic writing becomes”?

Quantitative comparison of degrees of coselectional constraint (Shadrova, 2020)

But what is coselectional constraint?

- is it: the strength of lexical association between words? I.e. “the more strongly individual words are associated, the more idiomatic writing becomes”? → statistical approach – but do words even have probabilities? (probably not)

Quantitative comparison of degrees of coselectional constraint (Shadrova, 2020)

But what is coselectional constraint?

- is it: the strength of lexical association between words? I.e. “the more strongly individual words are associated, the more idiomatic writing becomes”? → statistical approach – but do words even have probabilities? (probably not)
- or is it: an overarching measure of connectivity across the lexicon? I.e. “the tighter groups of words are bound while being less bound to other groups, the more idiomatic writing becomes”?

Quantitative comparison of degrees of coselectional constraint (Shadrova, 2020)

But what is coselectional constraint?

- is it: the strength of lexical association between words? I.e. “the more strongly individual words are associated, the more idiomatic writing becomes”? → statistical approach – but do words even have probabilities? (probably not)
- or is it: an overarching measure of connectivity across the lexicon? I.e. “the tighter groups of words are bound while being less bound to other groups, the more idiomatic writing becomes”? → graph-based approach – but how do aspects of the linguistic model correspond to aspects of graph theory?

Theses for discussion

Thesis 4: Different research questions require different operationalizations. For a precise quantitative analysis, we need to define the interfaces between the linguistic model, the quantitative model, and the data (the annotations)

Theses

1. Automatic analysis and NLP in its present state does not suffice for most linguistic research questions.
2. Often, our linguistic concepts are not well-defined and easy to operationalize – iterative annotation, especially of higher level concepts, *is* modeling.
3. Corpus linguistics needs to account for the high complexity, variability, and path-dependence of naturalistic linguistic data.
4. Quantitative analysis requires precise definitions of the mappings between the quantitative model, the linguistic model, and the data.

How do we tackle this complexity?

Different ways to divide and conquer:

- in-depth analysis on specialized, small to mid-sized corpora (SMISC, Lüdeling et al., 2021)
  - → requires quantitative frameworks that work on smaller data, such as graph metrics or Bayesian statistics, or that allow for a quantification of individual paths in path-dependent development (Dynamic Systems Theory)
- additive modeling to infer from surface phenomena to more abstract concepts
- leveraging applications from computational linguistics in more effective ways in linguistic research
  - → can the mechanics of blackbox computational models also tell us something new about language? Do surface-near NLP models have an equivalent in linguistic modeling? Or can they complement our day-to-day work?

Schedule

	Today
11:15–11:45	Introduction: Linguistic modeling and analysis
11:45–12:45	The group and the individual: Complementary dimensions of language development *Wander Lowie*
13:45–14:45	A comparison of frequentist and Baysian models of variation: The problems of priors and sample size *Natalia Levshina*
	Tomorrow
11:45–12:45	Corpora, inference, and models of register distribution *Felix Bildhauer, Elisabeth Pankratz, Roland Schäfer*
12:45–13:15	Deviation of proportions as the basis for a keyness measure *Christof Schöch, Julia Dudar, Cora Rok, Keli Du*
13:15–13:45	Machine Learning and syntactic theory: Focus on German and German varieties *Giuseppe Samo*
13:45–14:15	Discussion & Farewell

Speakers in a 60min slot are kindly asked to allow for 15min of discussion, in 30min slots please leave room for 10min.

If you are interested, please share your slides with us:
`dgfs2021.ccmlma@lists.hu-berlin.de`
and allow us to publish them on our workshop website:
`https://hu.berlin/dgfs2021-ccmlma`

Thank you for your attention!

anna.shadrova@hu-berlin.de
martin.klotz@hu-berlin.de
anke.luedeling@hu-berlin.de
https://hu.berlin/corpling

We gratefully acknowledge funding by the Deutsche Forschungsgemeinschaft for SFB 1412, 416591334 and RUEG, FOR 2537/1, 394844736

References I

Baayen, R. H. (2001). *Word Frequency Distributions*. Kluwer, Dordrecht / Boston / London.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krüger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language Models are Few-Shot Learners. *ArXiv*, abs/2005.14165.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Harbecke, D. (2021). Explaining Natural Language Processing Classifiers with Occlusion and Language Modeling.

Lüdeling, A., Hirschmann, H., Shadrova, A., and Wan, S. (2021). Tiefe Analyse von Lernerkorpora. Deutsch in Europa. Sprachpolitisch, grammatisch, methodisch., pages 235 – 283. de Gruyter, Berlin [u.a.].

Lukassek, J., Akbari, R., and Lüdeling, A. (2021). *Richtlinie zur morphologischen Annotation von Nomina in Falko*.

Mann, W. C. and Thompson, S. A. (1987). *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute Los Angeles.

Riester, A. and Baumann, S. (2017). The RefLex scheme-annotation guidelines.

Shadrova, A. (2020). *Measuring coselectional constraint in learner corpora: A graph-based approach*. Univ.-Dissertation, Humboldt-Universität zu Berlin.

Wan, S. (in prep.). *Argumentationsstrategien von chinesischen Deutschlernern*. Univ.Diss., Humboldt-Universität zu Berlin.

References II

Wiese, H., Alexiadou, A., Allen, S., Bunk, O., Gagarina, N., Iefremenko, K., Jahns, E., Klotz, M., Krause, T., Labrenz, A., Lüdeling, A., Martynova, M., Neuhaus, K., Pashkova, T., Rizou, V., Rosemarie, T., Schroeder, C., Szucsich, L., Tsehaye, W., Zerbian, S., and Zuban, Y. (2020). RUEG Corpus.

Zeldes, A. (2012). *Productivity in Argument Selection. From Morphology to Syntax.* Trends in Linguistics. De Gruyter Mouton, Berlin.

Discussion: Where we started

1. Automatic analysis and NLP in its present state does not suffice for most linguistic research questions.
2. Often, our linguistic concepts are not well-defined and easy to operationalize – iterative annotation, especially of higher level concepts, *is* modeling.
3. Corpus linguistics needs to account for the high complexity, variability, and path-dependence of naturalistic linguistic data.
4. Quantitative analysis requires precise definitions of the mappings between the quantitative model, the linguistic model, and the data.

Discussion: Remarks, problems, questions

- What is the idea (philosophy?) behind our selection of methods and behind a method itself? And are we always aware of model assumptions and do we respect them at all time? (1–4)
- Do we only gain from integrating new methods? (1–4)
- Sustainability: Does our work facilitate reproduction of results, can we easily instruct others in our approaches? (1–4)

Discussion: Remarks, problems, questions

- Can we (always) model language in a probabilistic framework? (1, 3, 4)
- How do we model w. r. t. the group vs. the individual, the outcome vs. the process? (2, 3, 4)
- We need complementing methods, not necessarily a replacement of different approaches and the integration of new methods (1)
- Is iterative, thorough modeling by annotation always worth the expense? (3, 4)