Approaches to linguistic annotation across languages and modalities

Volker Gast

Friedrich-Schiller-University of Jena

29 July, 2019

A rough schedule

1. Conceptual foundations of annotation
2. Annotation tools: Praat, ELAN, GraphAnno
3. Semantic and pragmatic annotation
   - Tense and modality
   - Scalar additive operators
   - Concessives
4. Multimodal annotation
   - Annotating gestures (in Idi)
   - Preprocessing and basic steps
   - Some (preliminary) results
5. Automatic annotation
   - Part of speech annotation, syntactic parsing
   - Named entities, coreference
   - Semantic and pragmatic annotation (Word2Vec, BERT)

What is linguistic data?

- Language as a **system** of symbols and rules (applying to these symbols), manifested in acoustic and visual signals.
- Linguistics as the study of (individual) linguistic systems and, more generally, the "human language faculty".
- An understanding of human language obviously requires the (comparative) study of various linguistic systems.
- Linguistic systems are instantiated in **linguistic matter**, i.e., physical manifestations of language, observable linguistic data.
- We can only observe linguistic matter, not linguistic systems; but unless we're a phonetician, we are primarily interested in the underlying systems.

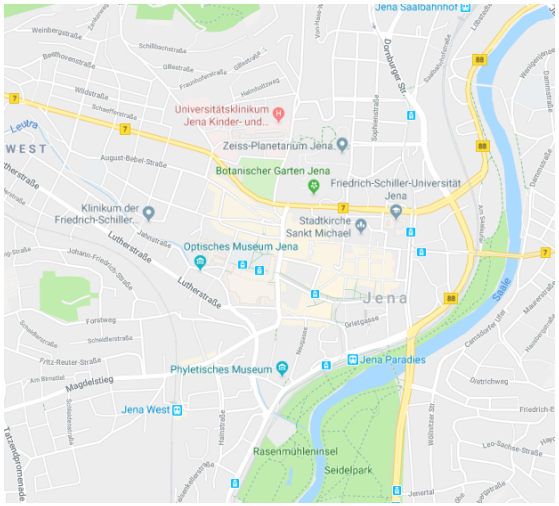
Perspectives on language

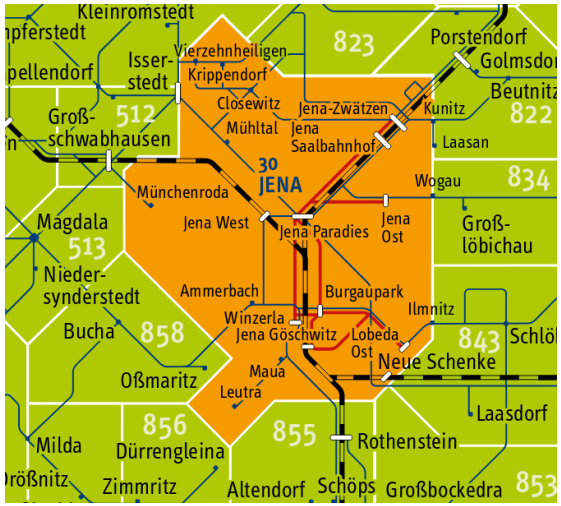
- Generative paradigm
  - Focus on the language production mechanisms, the 'generation' of linguistic matter by an underlying 'machinery'.
  - Central question: What are the design features of a (supposedly innate) generative grammar that produces linguistic matter?
- Usage-based paradigm
  - Focus on the linguistic matter itself
  - Central assumption: There is no major discrepancy between observable linguistic data and the 'mental representations' of this data.
  - Close relation to structuralism (without cognitivist background assumptions).
- Whatever perspective we take, we need to understand the structure observable in linguistic data.

Modelling linguistic data

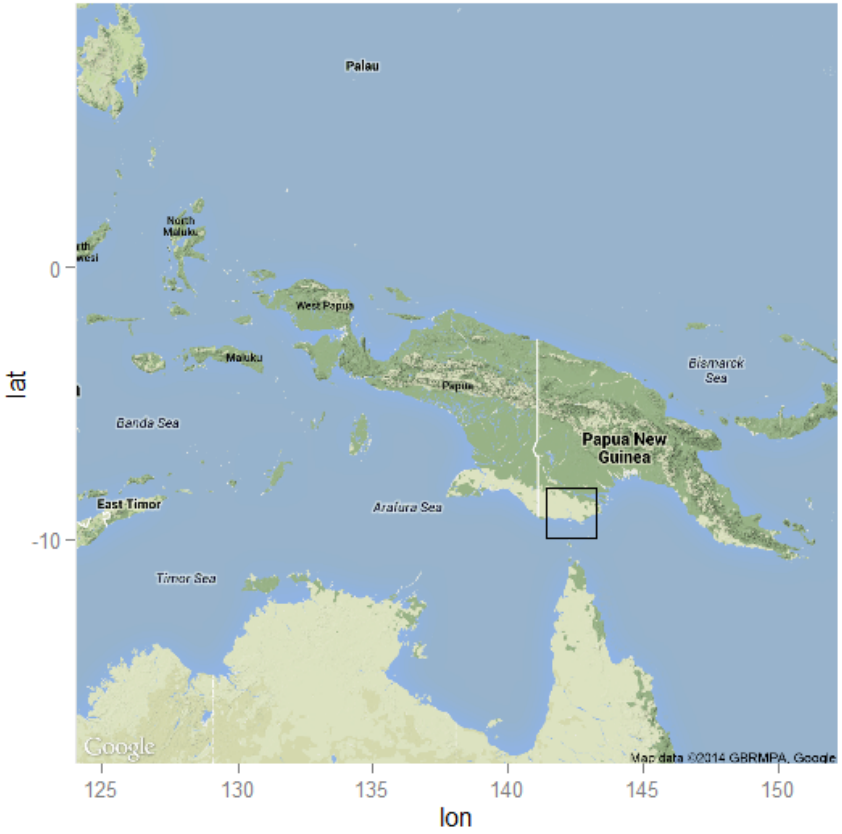
- Linguistic matter is ephemeral.
- If we want to represent linguistic systems, we have to create **models**, i.e. simplified representations of reality that reduce complexity and focus on those elements that we consider worthy of study.
- Modelling linguistic data makes it possible to represent it in a way that allows us to formulate generalizations, make predictions, provide explanations etc.

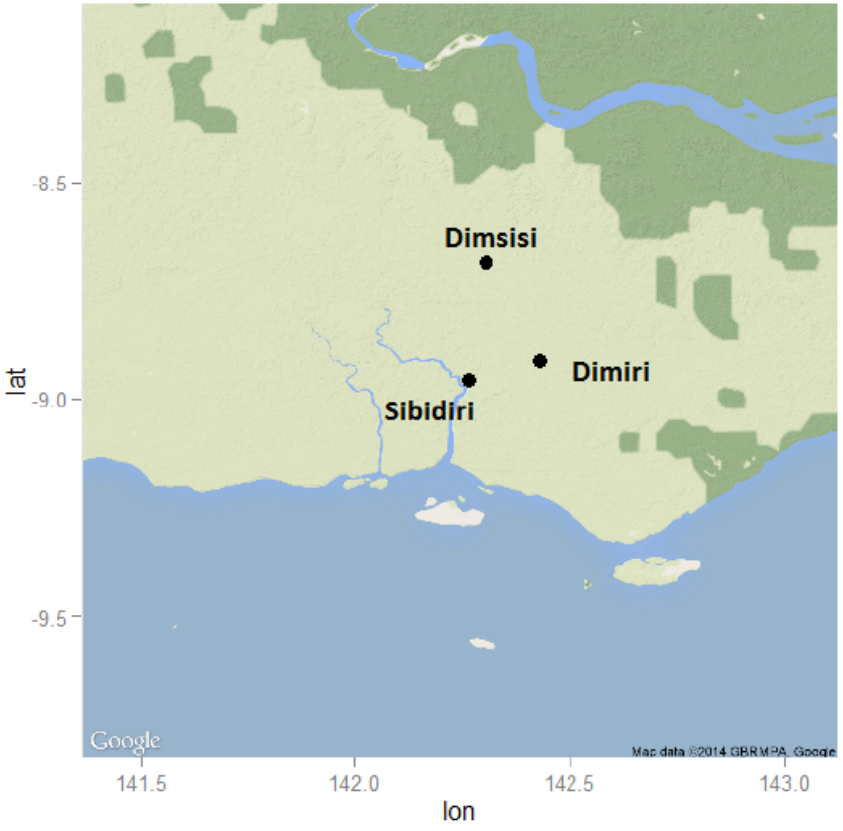
Models of Jena





Idi, a language of Southern PNG





Modelling phonetic matter

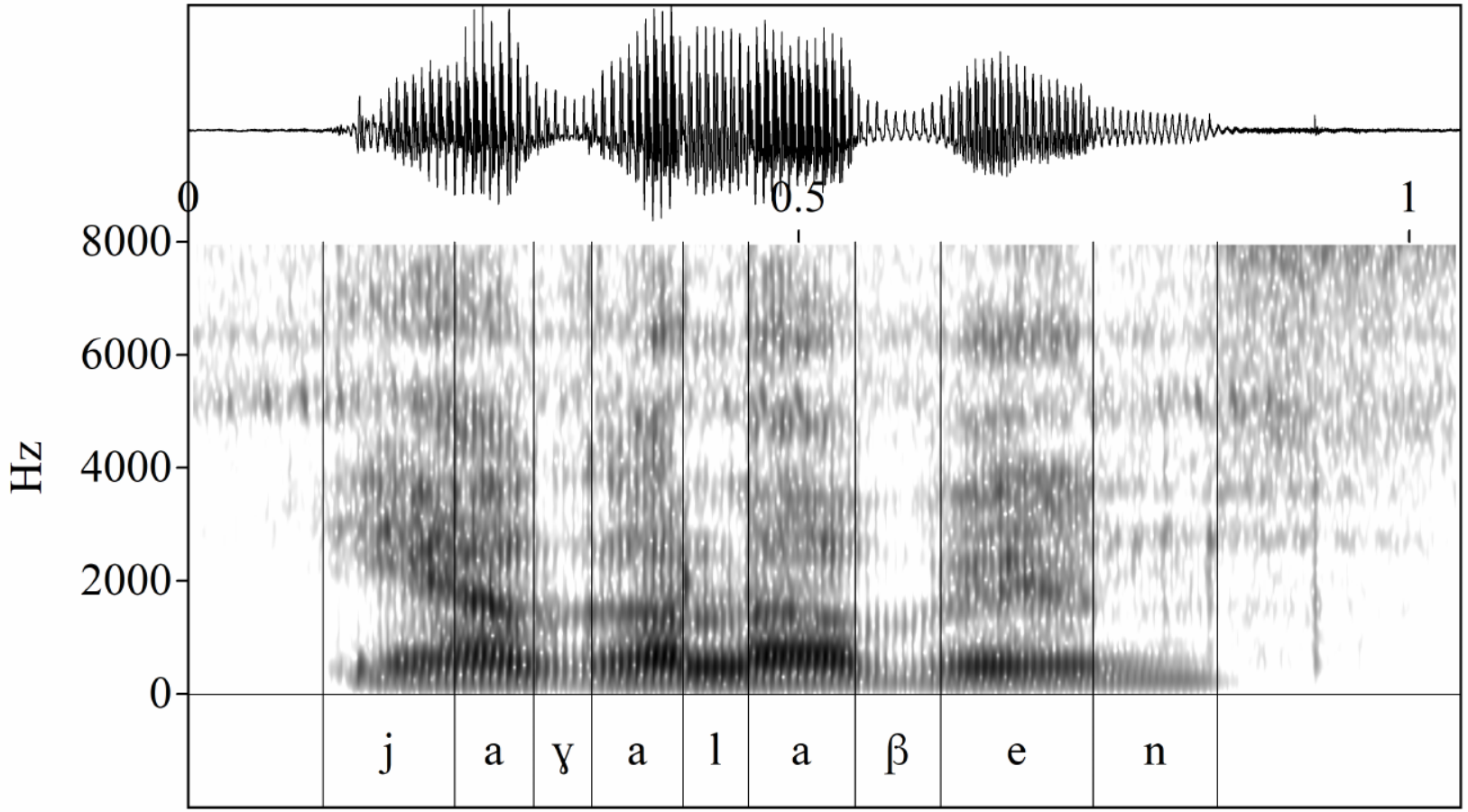
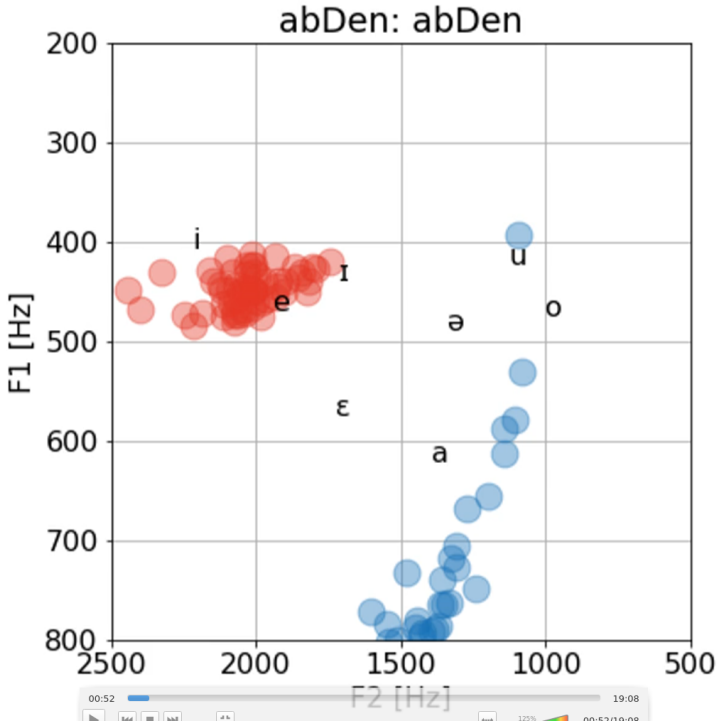


Figure: Notations of an acoustic signal

Representing language dynamically

- The dynamics of vowel articulation can sometimes better be represented in a dynamic way, e.g. in the form of video clips.



Structuralist modelling of linguistic data

- Phonetics relies on elements of physical modelling, e.g. by representing sound waves (air in motion) as mathematical functions.
- A phonetic transcription is a (linguistic) model of the (mathematical) model of sound waves; it classifies time spans into a finite inventory of sounds (phones), on the basis of physical similarity (e.g. F1 and F2 in the case of vowels).
- Phones can be regarded as sets of time spans with similar acoustic properties.
- Phones are further categorized into phonemes (sets of phones) – again, on the basis of similarity; but unlike in the classification of time spans into sounds, the distribution of the phones is taken into account.

Linguistic notations and annotations

- Once linguistic matter can be represented in terms of system elements, we can expand our model of description.
- The elementary symbols can be used to represent **linguistic objects** of different types.
- Linguistic objects are traditionally represented by delimiters of the type [...], /.../, {...}, etc.; they belong to different **layers** that are systematically related to each other.
- The structuralist model provides a system of notation.
- Essential tasks of linguistic analysis:
  - Assign **properties** to linguistic objects, i.e. classify them;
  - identify **relations** between linguistic objects.
- For instance, the linguistic object /yagalaben/ is of category V.
- The segment /galab/ is a verbal root, i.e. it is of category $\sqrt{V}$.
- Such attributions can be represented in the form of **annotations**.

The interlinear glossing system

- Linguistic typology is largely committed to an item-and-arrangement view of morphology, using interlinear glosses as a common annotation format.
- The item-and-arrangement view of morphology is simplistic in many cases.
- Alternative view: word-and-paradigm morphology, perhaps treating the morphological segments as expressing contraints.
- Each segment of the phonological layer is associated with the **constraints** specified in the morphological layer.

(1) [jagalaben]

/y-a-galab-e-n/
3Sg.O-Thm-open-RemPast-3Sg.A

'(S)he opened it.'

Annotation and language description

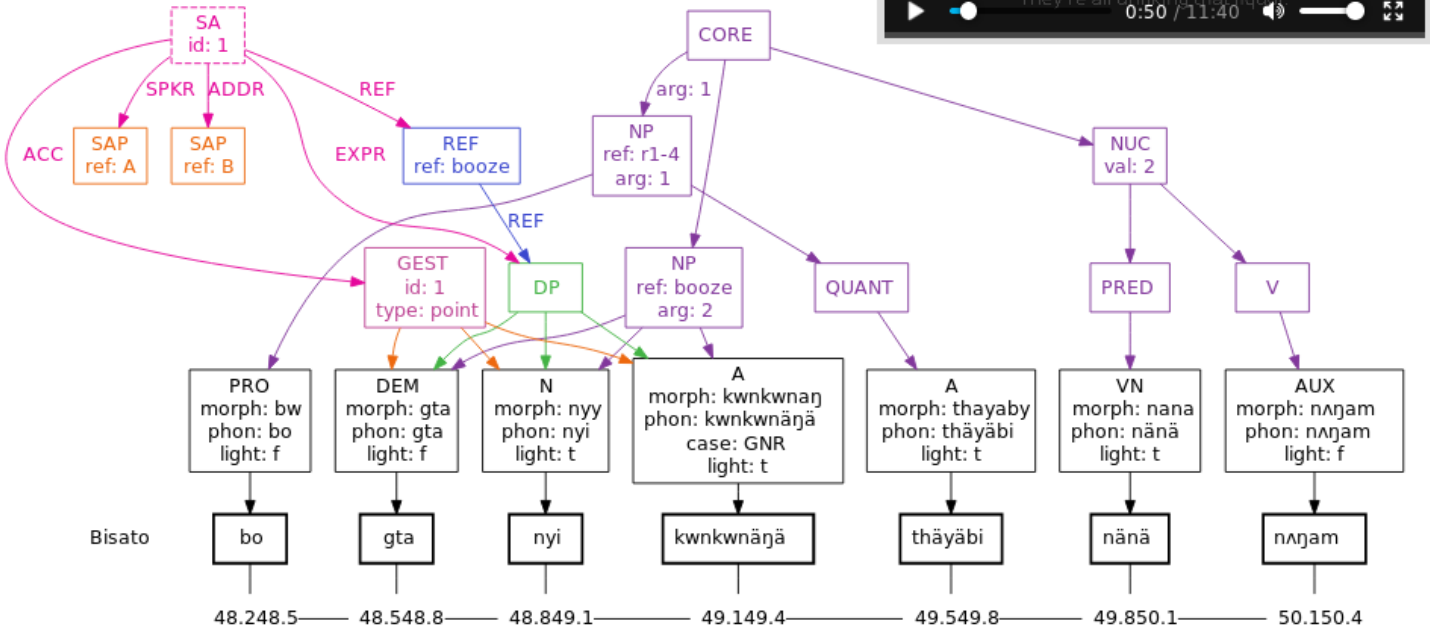
- In its most basic form, annotations are thus representations of properties, and relations between, linguistic objects.
- Differences to early structuralism:
  - Focus on observed language use
  - Computational treatment allowing automitization and quantitative analysis
  - Multi-level annotation: Linguistic objects can be classified along various dimensions, and these classifications are not necessarily hierarchically embedded.
  - Addition of further layers, e.g. semantic and pragmatic (linguistic and referential).
  - Multimodal annotations: assigning properties to gestures, signs etc.

The family problems picture task

- The Family Problems Picture Task,[1] a structured elicitation task.

[1] Lila San Roque et al. (2012). "Getting the Story Straight: Language Fieldwork Using a Narrative Problem-Solving Task". In: *Language Documentation & Conservation* 6, pp. 135–174.

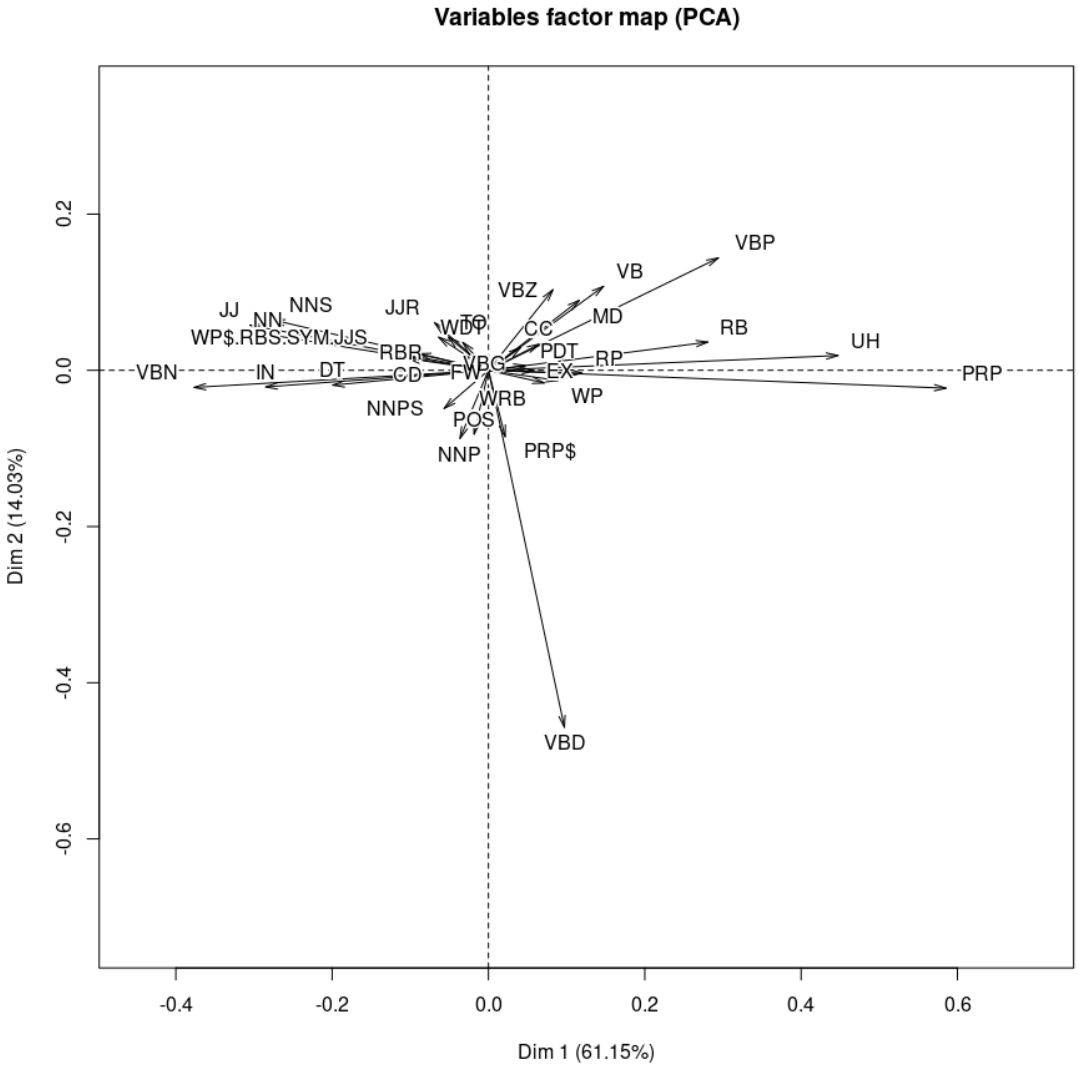
Gesture annotation



Morphological annotation: POS-tags

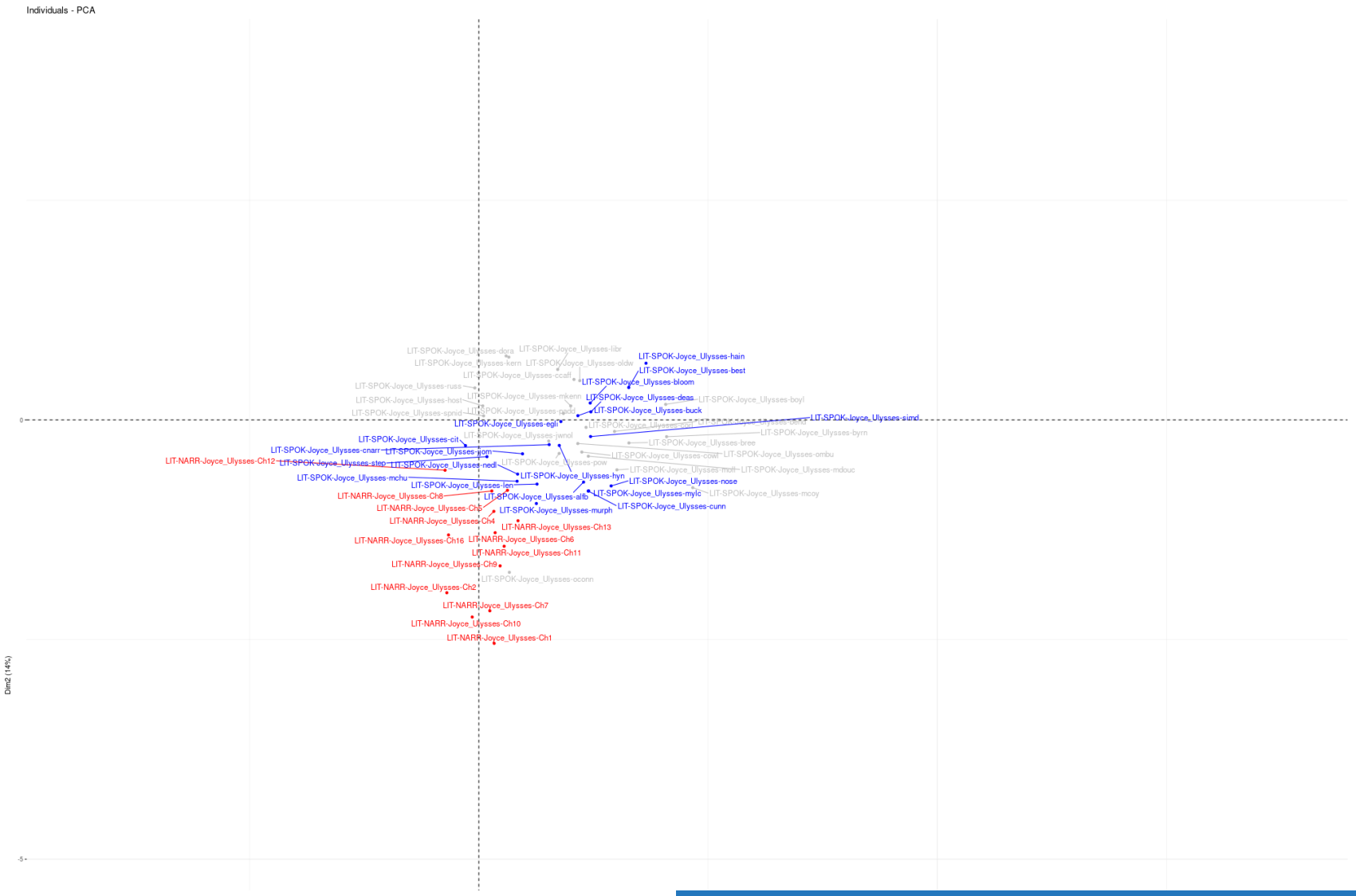
- Most traditional type of annotation, often used for quantitative analysis.
- Examples: Authorship attribution, speaker profiling
- Project: Comprehensive register studies of literary texts.
- Distributions of tags relative to specific speakers and narrators, with a focus on James Joyce.

Feature vectors



Distribution of ICE-IRL-Registers

Registers in Ulysses

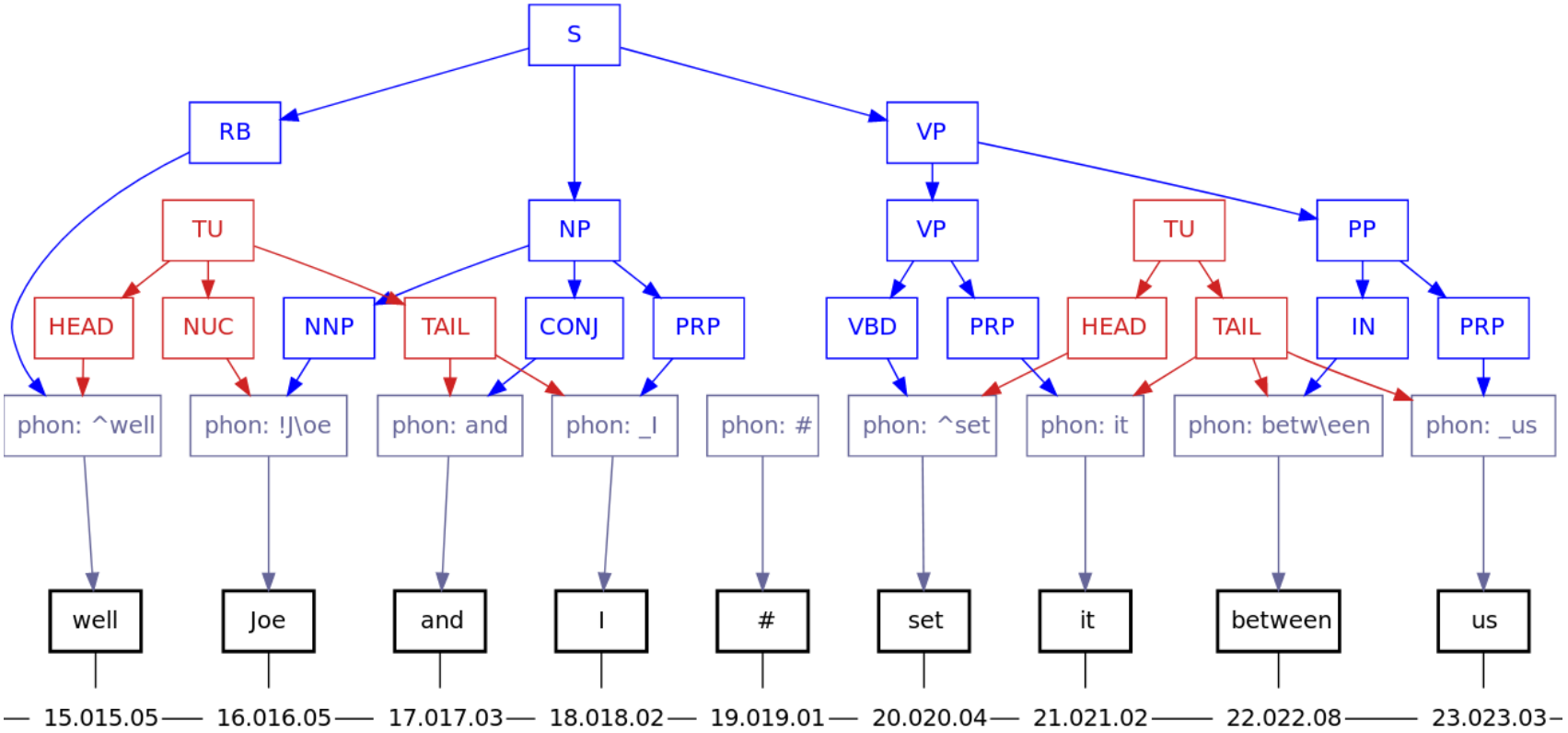


Registers in Hemingway

Phonological annotations: The LLC

```
1 1   1   10 1 1 B   11  ((of ^Spanish)) . graph\ology#
1 1   1   20 1 1 A   11  ^w=ell# .
1 1   1   30 1 1 A   11  ((if)) did ^y/ou _set _that# -
1 1   1   40 1 1 B   11  ^well !J\oe and _I#
1 1   1   50 1 1 B   11  ^set it betw\een _us#
1 1   1   60 1 1 B   11  ^actually !Joe 'set the :p\aper#
1 1   1   70 1 1 B   20  and *((3 to 4 sylls))*
1 1   1   80 1 1 A   11  *^w=ell# .
1 1   1   90 1 1 A   11  "^m/\ay* I _ask#
1 1   1  100 1 1 A   11  ^what goes !\into that paper n/ow#
1 1   1  110 1 1 A   11  be^cause I !have to adv=ise# .
1 1   1  120 1 1 A   21  ((a)) ^couple of people who are !d\oing [dhi: @]
1 1   1  130 1 1 B   11  well ^what you :d\/o#
1 1   1  140 1 2 B   12  ^is to - - ^this is sort of be:tween the :tw\/o of
1 1   1  140 1 1 B   12  _us#
1 1   1  150 1 1 B   11  ^what *you* :d\/o#
1 1   1  160 2 1 B   23  is to ^make sure that your 'own . !c\andidate
```

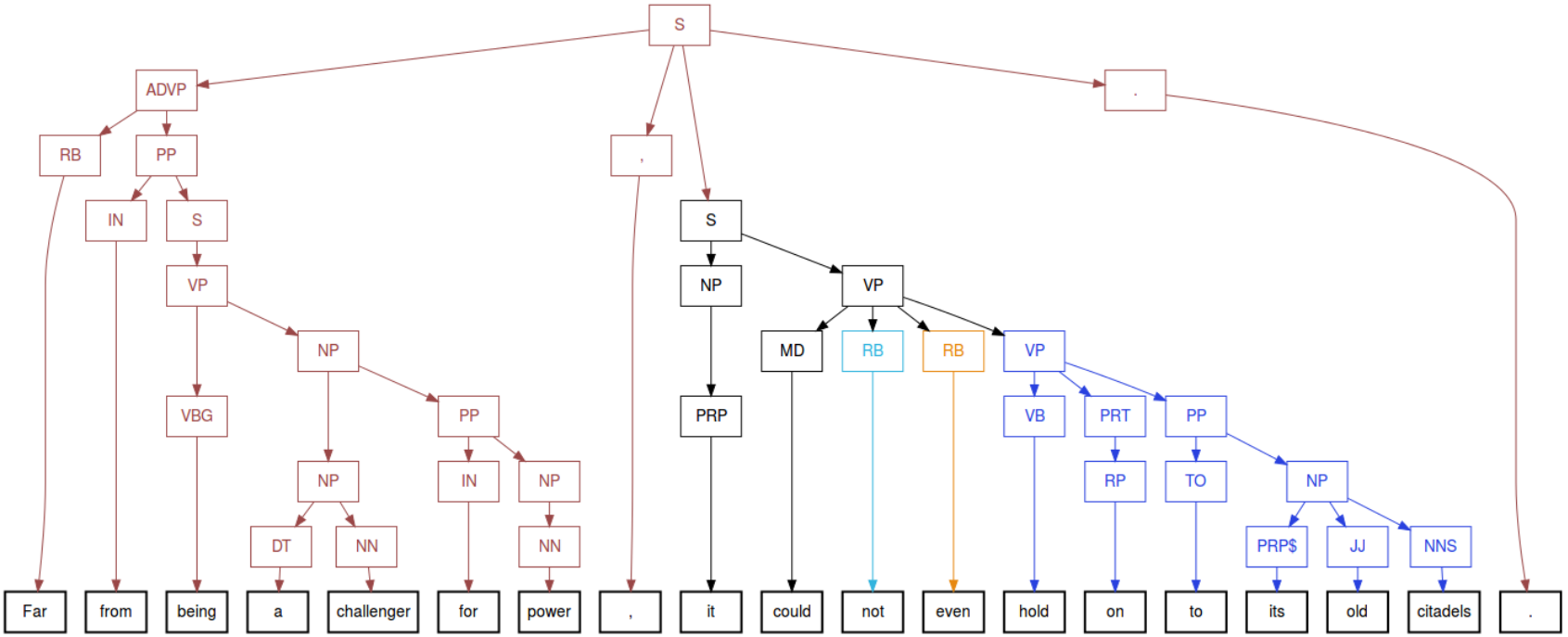
In GraphAnno . . .



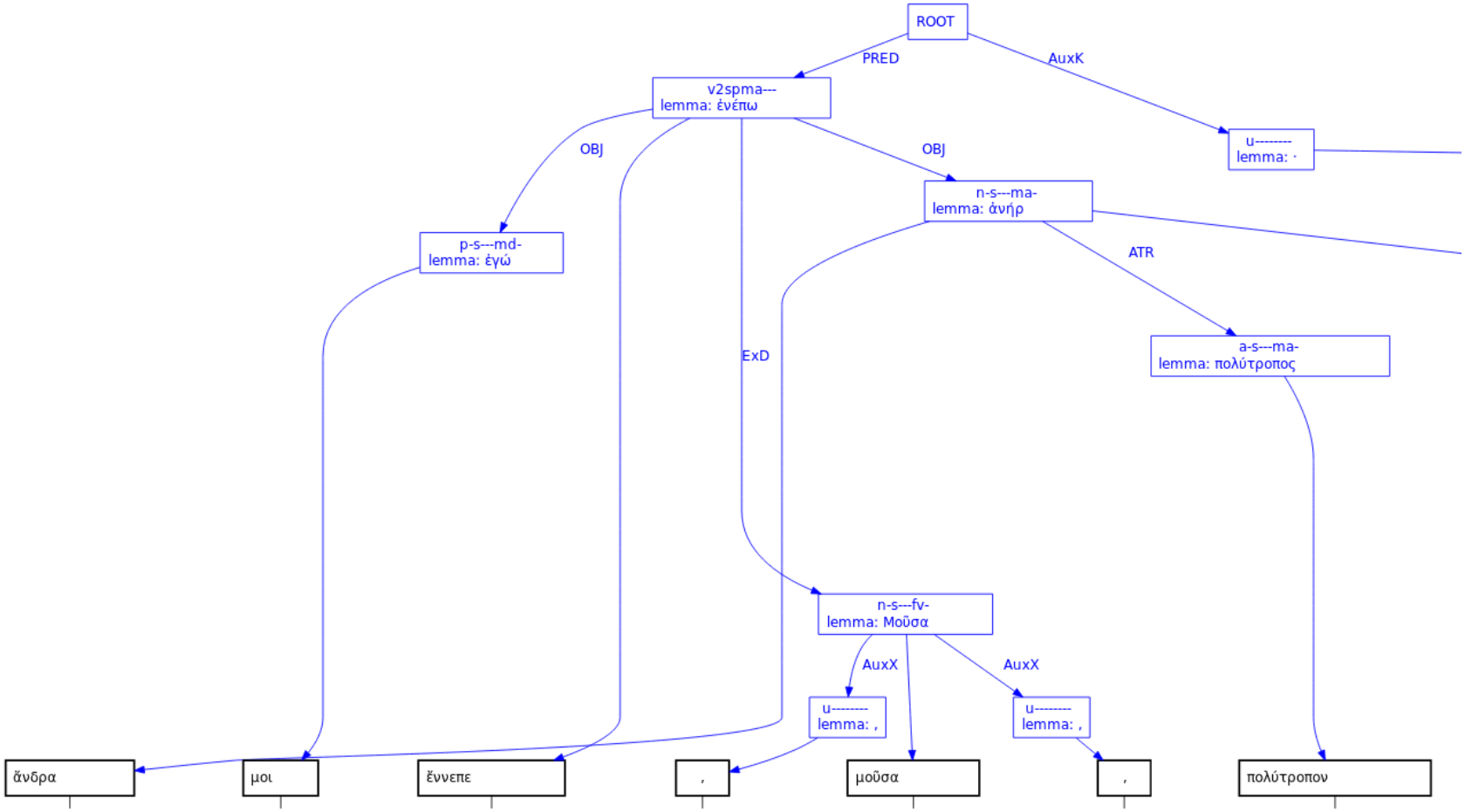
Syntactic annotations: Labelled bracketing

- At a certain time, NLP was concerned with phrase structure representations.
- These representations are not accepted within the generative community anymore.
- However, they can be useful to provide a certain scaffold for further annotations.

An example of phrase structure annotation



Syntactic annotations: Dependency relations



An example: The order of verb and subject in three languages

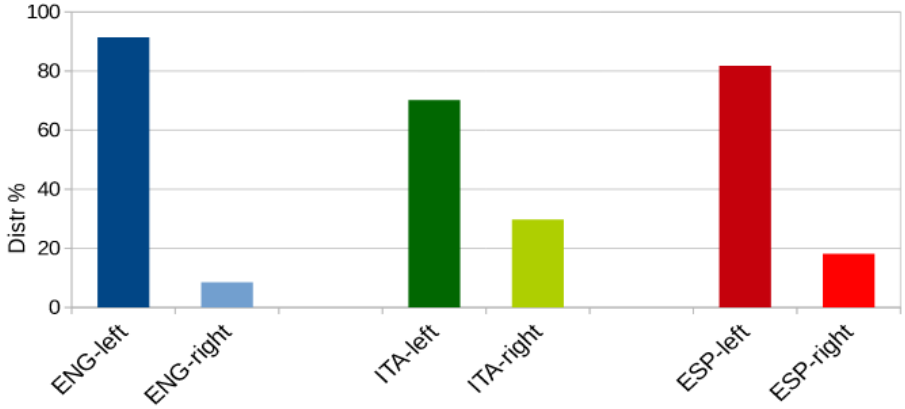


Figure 1: Distribution of right– vs left–headed non–pronominal *nsubj* relations in the three UD treebanks.

Figure: An example from Alzetta et al. 2018[2]

[2] Chiara Alzetta et al. (May 2018). "Universal Dependencies and Quantitative Typological Trends. A Case Study on Word Order". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. Miyazaki, Japan: European

Summary

- Annotation as a way of analysing language in a data-driven, but theoretically informed way.
- Particularly important for a usage-based framework, assuming that speakers store frequency information.
- Wide range of applicability, from language documentation (Idi) and quantitative typology to author/speaker profiling.