A Sentence is a Sentence is a Sentence? Parallels and Differences between the Segmentation of Oral and Historical Language Data

Zarah Weiss\textsuperscript{a}, Gohar Schnelle\textsuperscript{b}, Anke Lüdeling\textsuperscript{c} & Detmar Meurers\textsuperscript{d}

\textsuperscript{a}University of Tübingen; \textsuperscript{b}Humboldt Universität zu Berlin

A Sentence is a Sentence is a Sentence?

- Sentences are crucial units of linguistic analysis, but their definition is often biased by written language (Schmidt, 2016)
- Widespread notions confound well-formedness (syntactic) and graphematic markers (graphematic), e.g., Duden definition
- We present alternative segmentation units for non-standard data
- We demonstrate which segmentation units across non-standard language varieties share central requirements

Written Language Bias

- Implicit assumption of written language norm in linguistic theories (Hennig, 2008; Linell, 1982)
- Invalid for non-standard language systems such as spoken language, learner language, historical language, etc.
- Similar to comparative fallacy in SLA research (Bley-Vroman, 1983): analysis of I2 (only) in comparison to the target language
- Misrepresentation of systematicity of non-standard language system

Written Language Bias and NLP Tool Chains

Issue

- Sentence segmentation is a prerequisite for many NLP tasks
- Off-the-shelf sentence segmentizers heavily rely on punctuation
- This written language bias percolates down the NLP tool chain

The impact of missing punctuation on segmentation

- Data: Hamburg dependency treebank containing nearly 4 million tokens from www.helise.de (Foth, Kohn, Beuck & Menzel, 2014)
- NLP: spaCy for segmentation (Honnibal & Montani, 2017)
- Parse original tokens, tokens with 50\% less sentence-final punctuation, and tokens without any sentence-final punctuation

References


Segmentation of Historical and Oral German

ENHG t-Units (TU) in LangBank

- LangBank goal: create annotated corpus resource for ENHG (period without systematic punctuation (Hartweg & Wegera, 2005))
- t-Unit: “the shortest grammatically allowable sentence” (Hunt, 1965)
- “An (ENHG) t-unit consists of the head of a phrase and all of its arguments and adjuncts and nothing else” (Weiss & Schnelle, 2016)
- Based on notions from x-bar theory and basic topological field model
- Rules for atomism, discreteness, exhaustivity, continuity, and brevity
- High inter-rater agreement (Cohen’s $\kappa = 0.82$) between 3 annotators on 2,609 tokens from 5 texts written between 1532 and 1693

Maximal Syntactic Units (MSU) in SegCor

- SegCor goal: create segmentation for spoken German and French
- MSU = “maximal unit of related clauses” (Westpfahl & Gorisch, 2018); may be complete (simple or complex), abandoned, or non-sentential
- Based on the topological field model
- Rules for atomism, discreteness, exhaustivity, and discontinuity
- High inter-rater agreement (Cohen’s $\kappa = 0.69$) between 2 annotators on 22,150 tokens from 110 audio minutes (Westpfahl & Gorisch, 2018)

Historical and Spoken Language Phenomena

Historical Continuity from ENHG to Spoken German (Sandig, 1973)

1) das andere [...] hat Blättlein sind ein wenig rauber (Ridges corpus)
   the other one [...] has small. leaves are a bit rougher
2) and there he is on the roof was he (FOLK corpus; translation)

Non-Language Material

3) then they spoil the vine - (Ridges corpus; translation)
4) and then (laughter) (breathing) (FOLK corpus; translation)

Collaborative and Discontinuous Speech

T uhm (.) yes well now try to make a sentence
A if the whole thing is drawn the same size then (.) is (.) you go on you go on I made the start
B then you can see better #
A # what big what#
B # what is bigger and what is smaller
   translation of 7th graders in math class (Prediger & Wessel, 2018)

Criteria for Segmentation Units

Requirements for Segmentation Units by Auer (2010)

- Atomism: do not include other segments of their type
- Discreteness: do not allow overlapping segments
- Exhaustivity: capture all the linguistic material
- Coherence: use coherent linguistic descriptions

Additional Criteria

- Are reliable across and within annotators
- Avoid written language bias and comparative fallacy
- Can segments be discontinuous?
- Do segments need to come from one source?