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

Zarah Weiss

based on work done in collaboration with
Gohar Schnelle, Laura Perlitz, Carolin Odebrecht, Hagen Hirschmann,
Anke Lüdeling, and Detmar Meurers

“Unit segmentation in Spoken Interaction”
Segcor Workshop Orléans, June 3-5 2019
What is a Sentence?

- Sentence is a fundamental unit of linguistic analysis
- More than 200 sentence definitions (Eugen 1935)
- Naive understanding influenced by written language (Schmidt 2016; Bredel 2008, 2011)
- Assumes well-formedness (and graphematic markers)
- Confounds syntactic and graphematic notion
- Written language bias in linguistics (Linell 1982; Hennig 2008; Harris 1980)
A Sentence is a Sentence is a Sentence?

(1) It was dark outside. Everything was silent.
(2) It was dark outside, everything was silent.
(3) Olive oil, two shelves full.
According to *Duden Grammar* (Gallmann 2009, p. 763f)

1. Sentences are the maximal projection of a finite verb including all their arguments and adjuncts.
2. Sentences are syntactically well-formed units.
3. Sentences are the largest unit that can be derived by syntactically well-formed rules.
What is a Syntactic Sentence?

According to *Duden Grammar* (Gallmann 2009, p. 763f)

1. Sentences are the maximal projection of a finite verb including all their arguments and adjuncts.

2. Sentences are syntactically well-formed units.

3. Sentences are the largest unit that can be derived by syntactically well-formed rules.

→ Well-formedness in grammar refers to written language (Hennig 2009, 2008)
What is a Graphematic Sentence?

A sentence is a suprasegmental writing unit started by a sentence-initial majuscule and ended by a sentence-final punctuation mark. It does not contain these itself. The sentence-initial majuscule and the sentence-closing sign are terms that can be specified intragraphematically and are mutually dependent. Their alternation constitutes a sentence. (Schmidt 2016, p. 247)

- Follows punctuation theory Bredel (2008, 2011)
Written Language Bias

- Both definitions assume written language norm
- Example for written language bias in linguistics (Schmidt 2016; Hennig 2008; Linell 1982)
- Does not account for non-standard varieties of language: spoken language, learner language, etc.
- Relates to comparative fallacy in SLA
Comparative Fallacy

- Coined by Bley-Vroman (1983) for SLA
- Criticizes analysis of L2 in comparison to target language
  → Misrepresents particular systematicity of L2
- Linguistic analysis should expose linguistic properties
From Sentences to Sentential Units

Segmentation purpose

- Which properties should be exposed by an analysis?
  - Influences the decisions made in segmentation process

Sentential units (SU)

- Notion of ‘sentence’ is biased by theory and conventions
- May not be in interest of purpose-driven segmentation
- **Sentential unit** as maximal unit of related clauses
- **Keep clause** in its traditional sense as syntactic unit
Use Examples for Segmentation Units

- Study of syntactic phenomena and linguistic analysis
- Pragmatic analysis of discourse structures
- Sociolinguistic analysis of interlocutor interaction
- Unit for Natural Language Processing tool
- Unit for corpus representations (visualization, query)
Which Segmentation Do We Want?

T  Uhm (.) yes, well, now try to use the 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
T  Very good

7th graders in math class (Prediger & Wessel 2018)
Which Segmentation Do We Want?

T  Uhm (.) yes, well, now try to use the 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
T  Very good

7th graders in math class (Prediger & Wessel 2018)
Requirements of Sentential Segmentation Units

By Auer (2010)

1. **Exhaustivity**: Capture all the linguistic material
2. **Atomism**: Do not include other segments of their type
3. **Discreteness**: Do not allow overlapping segments
4. **Coherence on Linguistic Level**: Use linguistic descriptions

Additional criteria

- **Be reliably** applicable across and within annotators
- **Avoid written language bias** and comparative fallacy
Define Sentential Units Beyond A Written Norm

- Define sentential units for non-standard language
- Here for Early New High German (ENHG; 1482–1652)
- Bridge from ENHG to spoken language data

→ ENHG shares linguistic properties with spoken German
→ Historical continuity of non-normative syntactic patterns (Hennig 2007; Sandig 1973)
Dependent Clause with Verb-Last (VL)

(4) genannt Absinthium / welcher Name bis auf den
called Absinthium / which name until the
heutigen Tag in den Apotheken geblieben ist
present day in the pharmacies remained is
‘called Absinthium, which to the present day is the
name used used in pharmacies.’

Fuchs (1543)
weil-2 in Spoken German

(5)  a. weil ich Sie hierauf Aufmerksam machen will
     b. weil ich will Sie hierauf Aufmerksam machen
     c. ‘because I want to call your attention to this’

STUTTGART21 (2010)
(6) M. nennt sie *Rapiens Vitam / wann sie ist giftig*  
M. calls her Rapiens Vitam / because she is poisonous  
‘M. calls it *Rapiens Vitam* because it is poisonous.’

Tallat (1532)
The *LangBank* Project

- From 09/2015 to 09/2018
- Create annotated corpora for teaching and learning
- Languages: Classical Latin and ENHG
- Complement teaching with corpus-based work
- Investigate applicability of NLP for non-standard data
- ENHG Corpus available online¹

¹https://korpling.org/annis3/
Why Early New High German?

- Part of the German Arts and Literature curriculum
- Shares morpho-syntactic properties with contemporary German, e.g. verb asymmetry
- Interesting diachronic differences wrt. linguistic properties
  - Written ENHG register is forming
- Highly variable non-standard language variety
The *Ridges* Corpus

- Register in Diachronic German Science (RIDGES)
- Serves as empirical basis for project
- Corpus of herbal texts (1482 to 1914)
  (Lüdeling et al. 2018)
- LangBank time period: 1482 to 1652
- LangBank core: 14 books (80,095 *dipl* tokens)

→ Sentence ending punctuation emerge
  (Hartweg & Wegera 2005)
Punctuation Marks in ENHG (Bock, 1539)

nen man on merglichen nutz. Aber damit dem leser nichts abgehe/hab er sich beflissen/yedes kant/mit allen seinen geschlechten deutlich zubeschri
ben/vnd durch alle vmbstende vnd egentschaft/als an stengel/ziecke/
kant/blettern/blumen/safft/ farben vnd der gleichen von allen andern egentlichen zubehör/wo der fließig leser sein beschreibung wol bedenke vnd sie recht in die augen faffe /so habe er leicht zusehen/
welches kant solcher beschreibung zu sache oder nit. Wo aber yeman Lust hat zum gemelde/der möge zu diesem Büh/eyn ander treüiter büch kauf
senn/darim schöne figuren seint/als echtlich derzeit mit befindern fleiß aufs
gangen sein/vnd das bei nahe vmb so vil getts/als vmb unsers alleyn
erforderet wurde/wo es were mit figuren beladen/vi alsb also hatte er zu sein
lust vnd nutz ztwo bücher für eyns/ vnd die andern, so diese sein egent
Virgule (sentence final)
Punctuation Marks in ENHG (Bock, 1539)

Punctuation mark (sentence final)
Punctuation Marks in ENHG (cont.)

(7) das Wasser [...] braucht [...] Hieronymus von Braunschweig für das Abnehmen. Für den Braunschweig for the loosing weight. For the Hauptschwindel. Denen so Blut speien. dizziness. those who blood vomit ‘Hieronimus von Braunschweig uses this water against phthisis, dizziness, and to heal those, who vomit blood’

Megenberg (1482)
Sentential Units in *LangBank*

**Approach**
- No graphematic definition possible
- Syntactic definition required

**Purpose**
1. Annotation of syntactic phenomena
2. Basis for automatic and manual linguistic analysis
3. Unit for corpus representation (visualization, query)
4. (Investigation of diachronic changes from 1482 to 1914)
Assumptions from Contemporary German

Linguistic Units

- Words and Parts-of-Speech
- Phrases and constituents
- Clauses
- Sentential units based on original t-unit (Hunt 1965)
- ENHG t-unit, henceforth TU

Linguistic theories

- X-bar theory
- Topological field model
X-bar Theory

XP
  spec X’
  X’ adj
    X comp
      head

Introduction
Defining Sentences
Written Language Bias
Defining Sentential Units
Historical Continuity
The LangBank
T-Unit
Background
Annotation Principles
Basic Definition
Sentential Properties
Handling Ambiguities
Inter-Rater Reliability
Segmenting
Speech
The SegCor Project
Maximal Syntactic Unit
Generalizable Solutions
Open Issues
Conclusion
References
X-bar Theory (cont.)

```
VP
  NP  V'
    Julia  V'
        V  PP
        NP  at noon
          NP  Franca
            meets
```

Julia meets Franca at noon.
Segmenting Oral and Historical Language Data

Zarah Weiss

Introduction
Defining Sentences
Written Language Bias
Defining Sentential Units
Historical Continuity

The LangBank
T-Unit

Background
Annotation Principles
Basic Definition
Sentential Properties
Handling Ambiguities
Inter-Rater Reliability

Segmenting Speech
The SegCor Project
Maximal Syntactic Unit
Generalizable Solutions
Open Issues

Conclusion
References

X-bar Theory (cont.)

\[
\begin{align*}
\text{NP} & \quad \text{VP} \\
\text{Julia} & \quad \text{V'} \\
\text{VP} & \quad \text{PP} \\
\text{V} & \quad \text{NP} \\
\text{meets} & \quad \text{at noon} \\
\text{NP} & \quad \text{V'} \\
\text{Franca} & \text{meets} \\
\end{align*}
\]
Topological Field Model

- Non-hierarchical model of German clause structure
- Helps identify word order and constituency structure
- Arranged around German sentence bracket
- Model core applicable ENHG, but with limitations
- Wöllstein (2014); Pittner & Bermann (2004); Drach (1937)
### Segmenting Oral and Historical Language Data

Zarah Weiss

#### Introduction

**Defining Sentences**

**Written Language Bias**

**Defining Sentential Units**

**Historical Continuity**

#### The LangBank

**T-Unit**

**Background**

**Annotation Principles**

**Basic Definition**

**Sentential Properties**

**Handling Ambiguities**

**Inter-Rater Reliability**

#### Segmenting Speech

**The SegCor Project**

**Maximal Syntactic Unit**

**Generalizable Solutions**

**Open Issues**

#### Conclusion

#### References

---

**Topological Field Model (cont.)**

<table>
<thead>
<tr>
<th>N</th>
<th>Pre</th>
<th>LSB</th>
<th>Middle Field</th>
<th>RSB</th>
<th>Post</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_0$</td>
<td>Ich habe heute früher</td>
<td>gegessen</td>
<td>—</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

You have today earlier eaten
### Topological Field Model (cont.)

<table>
<thead>
<tr>
<th>N</th>
<th>Pre</th>
<th>LSB</th>
<th>Middle Field</th>
<th>RSB</th>
<th>Post</th>
</tr>
</thead>
<tbody>
<tr>
<td>S₀</td>
<td>Ich</td>
<td>habe</td>
<td>heute früher</td>
<td>gegessen</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>You</td>
<td>have</td>
<td>today earlier</td>
<td>eaten</td>
<td>—</td>
</tr>
<tr>
<td>S₂</td>
<td>—</td>
<td>Ist</td>
<td>das ein Satz?</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Is</td>
<td>this a sentence?</td>
<td></td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

---

**Example:**

You have today eaten — Is this a sentence? so that I can leave time.
### Topological Field Model (cont.)

<table>
<thead>
<tr>
<th>N</th>
<th>Pre</th>
<th>LSB</th>
<th>Middle Field</th>
<th>RSB</th>
<th>Post</th>
</tr>
</thead>
<tbody>
<tr>
<td>S₀</td>
<td>Ich</td>
<td>habe</td>
<td>heute früher</td>
<td>gegessen</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>You</td>
<td>have</td>
<td>today earlier</td>
<td>eaten</td>
<td></td>
</tr>
<tr>
<td>S₂</td>
<td>—</td>
<td>lst</td>
<td>das ein Satz?</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Is</td>
<td>this a sentence?</td>
<td>—</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>S₃</td>
<td>—</td>
<td>damit</td>
<td>ich mir Zeit</td>
<td>lassen kann</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>so.that</td>
<td></td>
<td>I me time</td>
<td>leave can</td>
<td>—</td>
</tr>
</tbody>
</table>

You have today earlier eaten — Is this a sentence? — so.that I me time leave can
### Segmenting Oral and Historical Language Data

Zarah Weiss

**Introduction**

- Defining Sentences
- Written Language Bias
- Defining Sentential Units
- Historical Continuity

### The LangBank T-Unit

- Background
- Annotation Principles
- Basic Definition
- Sentential Properties
- Handling Ambiguities
- Inter-Rater Reliability

### Segmenting Speech

- The SegCor Project
- Maximal Syntactic Unit
- Generalizable Solutions
- Open Issues

### Conclusion

### References

---

#### Topological Field Model (cont.)

<table>
<thead>
<tr>
<th>N</th>
<th>Pre</th>
<th>LSB</th>
<th>Middle Field</th>
<th>RSB</th>
<th>Post</th>
</tr>
</thead>
<tbody>
<tr>
<td>S₀</td>
<td>Ich</td>
<td>habe</td>
<td>heute früher</td>
<td>gegessen</td>
<td>S₃</td>
</tr>
<tr>
<td></td>
<td>You</td>
<td>have</td>
<td>today earlier</td>
<td>eaten</td>
<td></td>
</tr>
<tr>
<td>S₂</td>
<td>—</td>
<td>Ist</td>
<td>das ein Satz?</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Is</td>
<td>this a sentence?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S₃</td>
<td>—</td>
<td>damit</td>
<td>ich mir Zeit</td>
<td>lassen kann</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>so.that</td>
<td>I me time</td>
<td>leave can</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Starting with a T-unit

A t-unit is the “shortest grammatically allowable sentence into which (writing can be split) or minimally terminable unit” (Hunt 1965)
Starting with a T-unit

A t-unit is the “shortest grammatically allowable sentence into which (writing can be split) or minimally terminable unit” (Hunt 1965)

→ Circular if used as a sentence definition
Adjusting the Traditional T-unit

A TU consists of the head of a phrase and all of its arguments and adjuncts and nothing else. (Weiss & Schnelle 2016)
Adjusting the Traditional T-unit

A TU consists of the head of a phrase and all of its arguments and adjuncts and nothing else. (Weiss & Schnelle 2016)

- Does not require a predefined notion of a sentence
- Uses X-bar terminology to express “grammatically” and “minimally terminable”
- On its own misses crucial sentence properties
  → Elaborated on in the remaining rules
Atomism

Property

- SU do not contain other SU
- → they are atomic (Auer 2010)

Rule

- TU do not govern other TU
- The head of a TU may not be the argument or the adjunct of another head itself
Atomism in ENHG

(8) dass auch / hoch notwendig zu wissen [...] welche in that also / highly necessary to know [...] which in ihrem Leben / das ist / wann sie noch grün und their life / that is / when they still green and saftig sind ihre Tugend bald erzeigen juicy are their virtue soon show ‘that it is also necessary to know [...] which ones show their virtue while they are alive, i.e. while they are green and juicy’

von Bodenstein (1557)
Atomism in Speech

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 [...]  
B What is bigger and what is smaller

7th graders in math class (Prediger & Wessel 2018)
Discreteness

Property

- SU do not overlap with other SU
  → they are discrete (Auer 2010)

Rule

- TU may not overlap
- No phrase is part of more than one ENHG-TU.
Discreteness in ENHG

(9) das andere [...] / hat Blätten sind ein wenig rauher the other [...] / has small leaves are a bit rougher
‘The other one has leaves are a bit more rough.’

Fuchs (1543)
Discreteness in Speech

(10) and there he is on the roof was he

FOLK corpus²

²https://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscript&id=FOLK_E_00084_SE_01_T_01_DF_01&cID=c28&wID=c28
Exhaustivity

Property

- SU exhaustively partition a text
  → they are exhaustive (Auer 2010)

Rule

- A text has to be partitioned exhaustively into TU
- No material is left over
Exhaustivity in ENHG

(11) Da ich aber das Widerspiel erfahren habe / dass auf eine Zeit da der Samen unter dem / [...] / that at a time that the seeds under the / [...] Grund hervorgekommen ist / der über den carried ground emerge is / which over Winter grün verblieben ist / und nachwärts the winter green remained is / and im Sommer sehr groß geworden ist afterwards in the summer very big become is ‘but because I experienced the following counter-play: that after the seeds emerged from the earth [...] they remain green throughout winter and then grow up in summer’

Fuchs (1543)
Exhaustivity in Speech

A On the strip board [...] you can see it better because #
B # Because of Mehmet
A ((laughs)) # because it is in a row hence in a section #
B # a board #
A # or whatever you call it

7th graders in math class (Prediger & Wessel 2018)
(Dis)continuity

Question

- Is a sentential unit continuous or discontinuous?
- Depends on the purpose of the segmentation
- Here focus on syntactic analysis

Rule

- TU are continuous strings of tokens
- Exception: meta text artificially inserted by OCR
(Dis)continuity in ENHG

(12) dass auch / hoch  notwendig zu wissen ... welche in
that also / highly necessary to know ... which in
ihrem Leben / das ist / wann sie  noch grün und
their life / that is / when they still green and
saftig sind ihre  Tugend bald  erzeigen
juicy are their virtue soon show
‘that it is also necessary to know ... which ones show
their virtue while they are alive, i.e. while they are
green and juicy’

von Bodenstein (1557)
(Dis)continuity in Speech

A On the strip board [...] you can see it better because #
B # Because of Mehmet
A ((laughs)) # because it is in a row hence in a section #
B # a board #
A # or whatever you call it

7th graders in math class (Prediger & Wessel 2018)
Punctuation

Question

- Where is punctuation located in a SU?
- In practice, it is often SU-final
- Does not play a role for speech

Rule

- Punctuation is located at a TU’s outermost right periphery
Segmenting Oral and Historical Language Data
Zarah Weiss

Introduction
Defining Sentences
Written Language Bias
Defining Sentential Units
Historical Continuity
The LangBank
T-Unit
Background
Annotation Principles
Basic Definition
Sentential Properties
Handling Ambiguities
Inter-Rater Reliability
Segmenting Speech
The SegCor Project
Maximal Syntactic Unit
Generalizable Solutions
Open Issues
Conclusion
References

Punctuation in ENHG

(13) Das heimisch Eppich ist wohlschmeckend / aber es ist
the local ivy is tasty / but it is
dem Haupt böse
the jead evil
‘The local ivy is tasty. But it is bad for the head.’

von Megenberg (1482)
The local ivy is tasty. But it is bad for the head.'

von Megenberg (1482)
Handling Attachment Ambiguities

- Attachment ambiguities of phrases and clauses
  - Common in non-standard language, in particular ENHG
  - Challenges the uniqueness and discreteness of TU
TU Attachment Ambiguity

Ambiguity

► Is a unit a subordinate clause or a TU on its own?

Rule

► Minimize the maximal TU length in words
► Two shorter TU are preferred over one longer TU
→ Ambiguous cases are analyzed as TU
TU Attachment Ambiguity in ENHG

(15) Wermut ist ein Kraut mit vielen Zinken und Ästen / an welchem sind aschefarbene Blätter / on which/these are ash colored leaves
‘Vermouth is a herb with many branches ...’
A. ... on which there are ash colored leaves.’
B. ... On these there are ash colored leaves.’

Excerpt from Fuchs (1543)
TU Attachment Ambiguity in Speech

(16) Larry is grad fast ausm Fenster gefalln voll schlimm
Larry is just nearly out window fell really bad
A. ‘Larry just nearly fell out of the window really badly’
B. ‘Larry just nearly fell out of the window. So bad!’

FOLK corpus

---

3https://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscript&id=FOLK_E_00084_SE_01_T_01_DF_01%c1D=c15&wID=&textSize=200&contextSize=4
Phrase Attachment Ambiguity

Ambiguity

► To which TU does a phrase (or clause) belong?

Rule

► Minimize the maximal TU length in words
► Two shorter TU are preferred over one longer TU
→ Attachment to the shorter TU
→ Attachment to preceding TU for equal TU length
(17) So bedeuten nun diese zwei Blätter / das Gesetz in zwei Tafeln fest / von Gott gegeben / die anderen drei grünen Blättlein / die drei Personen dir zeigen ‘These two leaves symbolize the law in two tablets ... A. that were given by god. The other three green leaves B. The other three green leaves given by god ... show you three people’

Rosbachs (1588)
(18) So bedeuten nun diese zwei Blätter / das Gesetz in zwei Tafeln fest / von Gott gegeben / die anderen drei zwei tablets firm / by god given / the other three grünen Blättlein / die drei Personen dir zeigen green leaves / which three people you show ‘These two leaves symbolize the law in two tablets ... A. that were given by god. The other three green leaves B. The other three green leaves given by god ... show you three people’

Rosbachs (1588)
(19) So bedeuten nun diese zwei Blätter / das Gesetz in zwei Tafeln fest / von Gott gegeben / die anderen drei zwei tablets firm / by god given / the other three grünen Blättlein / die drei Personen dir zeigen green leaves / which three people you show ‘These two leaves symbolize the law in two tablets ... A. that were given by god. The other three green leaves B. The other three green leaves given by god ... show you three people’

Rosbachs (1588)
Phrase Attachment Ambiguity in Speech

(20) ihr könnt euch das mal gegenseitig [...] erklären was
you can you this now each other [...] explain what
ihr da gemacht habt am besten ihr habt das ja
you there dine have the best you have this yes
alle richtig gemacht
all correctly done
A. ‘It would be best you explain to each other what
you did. All of you did everything correct anyway.’
B. ‘Now explain to each other what you did. Ideally, all
of you did everything correct anyway.’

Prediger & Wessel (2018)
Phrase Attachment Ambiguity in Speech

(21) ihr könnt euch das mal gegenseitig [...] erklären was you can you this now each other [...] explain what ihr da gemacht habt am besten ihr habt das ja you there dine have the best you have this yes alle richtig gemacht all correctly done

A. ‘It would be best you explain to each other what you did. All of you did everything correct anyway.’
B. ‘Now explain to each other what you did. Ideally, all of you did everything correct anyway.’

Prediger & Wessel (2018)
(22) ihr könnt euch das mal gegenseitig [...] erklären was you can you this now each other [...] explain what ihr da gemacht habt am besten ihr habt das ja you there done you have the best you have this yes alle richtig gemacht all correctly done
A. ‘It would be best you explain to each other what you did. All of you did everything correct anyway.’
B. ‘Now explain to each other what you did. Ideally, all of you did everything correct anyway.’

Prediger & Wessel (2018)
Well-Formedness vs. Brevity

Issue

- SU in non-standard language may (not) be well-formed
- What to do if a well-formed analysis is possible?
  → Ambiguities between well-formed and short TU arise

Rule

- When in doubt, prefer well-formedness over brevity
Well-Formed vs. Brevity in ENHG

(23) aber sie trocknen die Zungen [...] wenn man Köl-
but they dry the tongues [...] if you köl-
und Haselbaum pflanzt zu Weinrebenwurzeln /
and hazel tree plant next to grape vine roots /
sō verderben sie die Reben
then spoil they the grape vine
‘But they dry the tongues [...]. If you plant köl tree and hazel tree next to grape vine roots, then they spoil the grape vine.’

von Megenberg (1482)
(24) I have already all the time been hanging in the no not all the time a minute I have been hanging in the line it’s quite funny when you don’t hear anything

FOLK corpus⁴

⁴https://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscript&id=FOLK_E_00084_SE_01_T_01_DF_01&cID=c10&wID=c10
Our TU Definition in one Slide

1. A TU consists of a phrasal head and all its arguments and adjuncts
2. TUs are independent, unique, exhaustive
3. TUs are continuous and do not start with punctuation
4. Ambiguous clause and phrase attachment is resolved by preference of brevity
5. Well-formedness trumps brevity
Inter-Rater Reliability

- End of every word is sentence boundary candidate
  - Binary ± sentence boundary annotation
    - 3 annotators
    - 5 text excerpts from 1532 to 1639
    - 2,609 tokens → ca. 5% sentence boundaries
  - Cohen’s Kappa ($\kappa$) = .82
  - Almost perfect agreement: $\kappa \geq .80$ (Landis & Koch 1977)
The SegCor Project

- Segmentation of Oral Corpora (06/16 to 02/19)
- Develop methods for segmentation of spoken language
- Languages: Spoken French and German
- Corpora: FOLK (German), CLAPI and ESLO (French)
- Westpfahl et al. (2018); Westpfahl & Gorisch (2018); Schmidt & Westpfahl (2018)
SegCor Guidelines: Purpose and Means

- Based on syntactic criteria
- Find phenomena typical for spoken language
- Offer information about syntactic phenomena
- Practical approach
- See Westpfahl et al. (2018, p. 3)
Segmenting Oral and Historical Language Data
Zarah Weiss

Introduction
Defining Sentences
Written Language Bias
Defining Sentential Units
Historical Continuity
The LangBank
T-Unit
Background
Annotation Principles
Basic Definition
Sentential Properties
Handling Ambiguities
Inter-Rater Reliability
Segmenting Speech
The SegCor Project
Maximal Syntactic Unit
Generalizable Solutions
Open Issues
Conclusion
References

Syntactic Segmentation Layers

- Topological fields
- Clause type (based on verb position)
- Maximal syntactic unit (MSU)
- (Certain speech phenomena)

→ Hierarchical multi-layer annotation approach
→ MSU serve as ‘maximal unit of related clauses’ (SU)
→ Based on same linguistic units and theories as TU
Types of Maximal Syntactic Units

- Simple sentential unit (complete, 1 clause)
- Complex sentential unit (complete, 2+ clauses)
- Abandoned unit (incomplete clause)
- Non-sentential unit (independent phrases, non-verbal)
Types of Maximal Syntactic Units

- Simple sentential unit (complete, 1 clause)
- Complex sentential unit (complete, 2+ clauses)
- Abandoned unit (incomplete clause)
- Non-sentential unit (independent phrases, non-verbal)

→ Currently no similar differentiation for TU
→ Applicable to current TU definition
Similarities between MSU and TU

- Maximal units of connected clauses (syntactic definition)
- Atomic, discrete, exhaustive, and coherent (Auer 2010)
- Differ in linguistic theories used to express concept
  - TU maximal phrasal projection
  - MSU all related clauses or independent/abandoned units
Shared phenomena

- Dialect and non-standard orthography
- Parentheses and ellipses, apo koinu
- Non-sentential and abandoned language material
- Syntactic similarities between ENHG and spoken German

Different challenges and cues

- Speech: collaborative or interrupted speech, pauses
- ENHG: no native speakers’ judgments, sections
Similar Solutions to Similar Challenges

Apo koinu

(25) and there he is on the roof was he (FOLK corpus)

Parentheses and insertions

(26) a. I’ve been waiting all the time in the no not all the time for a minute in the line (FOLK corpus)

b. because (.) madam minister (.) you don’t only know the conservatory (FOLK corpus)
Similar Solutions to Similar Challenges (cont.)

Free noun phrase
(27) a reasonable cat . the big black one (FOLK corpus)

Non-language material
(28) and then ((laughter)) ((breathing)) (FOLK corpus)
(29) then they spoil the vine - ¶ (von Megenberg, 1482)
Discontinuous Speech ↔ Continuity

T Why is it difficult to formulate?
A On the strip board, on the strip board you can see it better because #
B # Because of Mehmet
A ((laughs)) # because it is in a row hence in a section #
B # a board #
A # or whatever you call it
T Uhm what do you mean #

7th graders in math class (Prediger & Wessel 2018)
Discontinuous Speech \& Continuity

T Why is it difficult to formulate?
A On the strip board, on the strip board you can see it better because #
B # Because of Mehmet
A ((laughs)) # because it is in a row hence in a section #
B # a board #
A # or whatever you call it
T Uhm what do you mean #

7th graders in math class (Prediger & Wessel 2018)
Collaborative Speech & Continuity

T  Uhm (.) yes, well, now try to use the 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
T  Very good

7th graders in math class (Prediger & Wessel 2018)
Collaborative Speech § Continuity

T  Uhmm (.) yes, well, now try to use the 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
T  Very good

7th graders in math class (Prediger & Wessel 2018)
Conclusion

Sentence segmentation

- Sentence definitions suffer from written language bias
- Different purposes call for different SU
- Yet many principles can be shared across purposes
- Only few phenomena require specialized rules

→ Multi-layer annotation of SU (see SegCor annotation)
Conclusion (cont.)

Generalizability of TU and MSU

- Share assumptions and principles
- Often yield same or comparable segmentation
- TU lack applicability to collaborative speech
- MSU relies on topological fields and native-speaker intuition → limits generalizability
Outlook

- COLD project: 04/2019 to 04/2021
  - Investigate interaction of teachers and students
  - Adaptation strategies of teachers to language learners
  - Challenges
    - How do we segment our classroom interactions?
    - Multiple interlocutors and parallel discourses
    - Incomplete and collaborative utterances
  - Currently: decide transcription and segmentation

5https://www.die-bonn.de/COLD/
References


Segmenting Oral and Historical Language Data
Zarah Weiss

Introduction
Defining Sentences
Written Language Bias
Defining Sentential Units
Historical Continuity

The LangBank
T-Unit

Background
Annotation Principles
Basic Definition
Sentential Properties
Handling Ambiguities
Inter-Rater Reliability

Segmenting Speech

The SegCor Project
Maximal Syntactic Unit
Generalizable Solutions
Open Issues

Conclusion

References


Segmenting Oral and Historical Language Data
Zarah Weiss

Introduction
Defining Sentences
Written Language Bias
Defining Sentential Units
Historical Continuity

The LangBank
T-Unit
Background
Annotation Principles
Basic Definition
Sentential Properties
Handling Ambiguities
Inter-Rater Reliability

Segmenting Speech
The SegCor Project
Maximal Syntactic Unit
Generalizable Solutions
Open Issues

Conclusion

References


STUTTGART21 (2010). STUTTGART21: German panel discussion. Phoenix TV.


Segmenting Oral and Historical Language Data
Zarah Weiss

Introduction
Defining Sentences
Written Language Bias
Defining Sentential Units
Historical Continuity

The LangBank
T-Unit
Background
Annotation Principles
Basic Definition
Sentential Properties
Handling Ambiguities
Inter-Rater Reliability

Segmenting Speech
The SegCor Project
Maximal Syntactic Unit
Generalizable Solutions
Open Issues

Conclusion

References
Written Language Bias and Comparative Fallacy

The learner’s system is worthy of study in its own right, not just as a degenerate form of the target system Bley-Vroman (1983, p. 4)

In grammar description, linguists tend to regard the peculiarities of the former [= spoken language] as ‘deviations’ rather than as independent structural principles. (translated from Small 1985a: 13)
(30) Die Wasser [...] sind gut / jedoch nicht so stark und kräftig / als wie durch das Instrument destilliert vigorous / than as through the instrument distilled ‘The waters are good but not as strong and vigorous as waters distilled with the instrument.’

Libavius (1603)
(31) Eine Eisenbahnfahrt im Tiefbahnhof findet unter keinen anderen eisenbahntechnischen Voraussetzungen statt als wie im Kopfbahnhof. ‘A railway journey in an underground station does not take place under any other technical railway conditions than in a terminus station.’

STUTTGART21 (2010)
Another Punctuation Example

(32) das is mehr für gelehrt Leute / als die sich
this is more for educated people / as those themselves
mit dem Werk zu belustigen begehre. In welchem er
with the work to entertain seek. in which he
aus Plinius sehr viel genommen
from Plinius very much took
‘this is rather for researchers than for those who read
for pleasure the book in which he adopted a lot from
Plinius’

Rhagor (1693c)
(33) derer man alle morgens und abends zwei Skrupel dem Schwindsüchtigen [soll geben] / scrupuli the tuberculosis infected shall give / Adam von Bodenstein (1557)
Another TU Attachment Ambiguity in ENHG

(34) aber von dem Baum der Erkenntnis Gutes und Böses
but of the tree of knowledge good and bad
sollst du nicht essen / denn welches Tages du
shall you not eat / because which day you
davon isst / sollst du des Todes sterben
of it eat / shall you the death die
‘but of the tree of knowledge about good and evil you
shall not eat, because on the day that you eat of it,
you will die’

Rosbachs (1588)
(35) aber von dem Baum der Erkenntnis Gutes und Böses
sollst du nicht essen / denn welches Tages du
shall you not eat / because which day you
davon isst / sollst du des Todes sterben
of it eat / shall you the death die
‘but of the tree of knowledge about good and evil you
shall not eat, because on the day that you eat of it,
you will die’

Rosbachs (1588)
A Sentence is a Sentence is a Sentence?

DOGS SHEEP" from Feilke (2010)
Spoken Learner Language

A which which what is your opinion?
B (1.0) maybe er (5.0) he (7.0)
A long time? or it’s for for you it’s a major mistake or a small mistake?
B maybe three months
A three months for this one okay for me it’s ten
B ten?
A ten years
B yeah ten years oh very long

L2 interaction on prison sentences (Foster & Skehan 1996)
Interrupted Turns L1 German

T Why is it difficult to formulate?
A On the strip board, on the strip board you can see it better because #
B # Because of Mehmet
A ((laughs)) # because it is in a row hence in a section #
B # a board #
A # or whatever you call it
T Uhm what do you mean #

7th graders in math class (Prediger & Wessel 2018)
Interrupted Turns L1 German

T  Why is it difficult to formulate?
A  On the strip board, on the strip board you can see it better because #
B  # Because of Mehmet
A  ((laughs)) # because it is in a row hence in a section #
B  # a board #
A  # or whatever you call it
T  Uhm what do you mean #

7th graders in math class (Prediger & Wessel 2018)
Interrupted Turns L1 German (cont.)

T  Uhm what do you mean #
C  # a line #
T  # a stripe?
B  Because #
A  # That #
B  # Mehmet is there
A  ((laughs))
T  No speak again constructively

7th graders in math class (Prediger & Wessel 2018)
Interrupted Turns L1 German (cont.)

T  Uhm what do you mean #
C  # a line #
T  # a stripe?
B  Because #
A  # That #
B  # Mehmet is there
A  ((laughs))
T  No speak again constructively

7th graders in math class (Prediger & Wessel 2018)
Interrupted Turns L1 German (cont.)

T  Uhm what do you mean 
C  # a line 
T  # a stripe? 
B  Because 
A  # That 
B  # Mehmet is there 
A  ((laughs)) 
T  No speak again constructively 

7th graders in math class (Prediger & Wessel 2018)
Incomplete MSU

FR  he’s just a little bit of a klutz
EG  yes yes exactly (.) that’s the problem (.) Pepper obviously didn’t go out either
FR  yes
EG  and only Larry does such a thing here
FR  a reasonable cat (.) the big black one
EG  ((laughing)) yeah kind of
Incomplete MSU

FR  he’s just a little bit of a klutz

EG  yes yes exactly (.) that’s the problem (.) Pepper obviously didn’t go out either

FR  yes

EG  and only Larry does such a thing here

FR  a reasonable cat (.) the big black one

EG  ((laughing)) yeah kind of

- MSU annotation as non-sentential unit (Westpfahl et al. 2018, p. 39)
Excerpt from von Bodenstein (1557)

(36) [...] von dem L. C. meldet / dass er der erste
[...] of whom L. C. reports / that he the first
gewesen ist / so den Feldbau die lateinische
been has / who the agriculture the Latin
Sprache gelehrt hat
language taught has
‘of whom L. C. reports that he has been the first who
brought Latin into agriculture’
(37) doch mögen wir [...] dafür gebrauchen das Kraut
but want we [...] for this use the herb
so man Welsamen nennt / denn es der Kraft
which one Welsamen calls / because it the power
[...] nach dem rechten Seriphium ganz
[...] following the right Seriphium completely
gleich ist
similar is
‘but for this we want to use the herb that is called
Welsamen because it is very similar to the right
Seriphium in terms of its powers’

Excerpt from Fuchs (1543)
Historical Continuity of Apo Koinu

(38) Uns ist in alten mæren wunders vil geseit / von Us is in old tales wonderful much told / of helden lobebären, von grôzer arebeit, / von freuden, heroes laudable of great labor / of joys hôchgezîten, von weinen und von klagen, / von celebrations of crying and laments / of bold küener recken strîten muget ir nu wunder warriors’ fights may you now wonderful hear hœren sagen.
told
‘In wonderful stories we have heard of laudable heroes, great labor, joy, celebrations, crying and laments, and fighting warriors may you now hear wonderful tales’
Heavy Right Periphery in Spoken German

(39) Ich zeige es Ihnen deshalb, weil ich will Sie aufmerksam machen, auf hier diesen Abschnitt hier. ‘I am showing you this, because I want you to make attention to this paragraph here’

STUTTGART21 (2010)
(40)  darin der Samen frei verborgen [liegt] eine lange Zeit
in this the seed free hidden lies a long time
‘in this lies the seed openly hidden for a long time’

Rosbachs (1588)