Zarah Weiß & Gohar Schnelle

Early New High German
Sentence Segmentation

Annotation Guidelines

Version 4.0

Contents

1 Introduction

The following guidelines were designed to allow for a consistent sentence segmentation of Early New High German (ENHG) texts. As these texts include only partial or ambiguous punctuation, a mainly graphematic sentence definition is not applicable. This challenges automatic and manual sentence segmentation approaches alike: According to Schmidt (2016, p. 216) over 200 non-graphematic sentence definitions have been suggested by linguists so far.[1] Yet, proper sentence segmentation is an important basis for further linguistic annotations.

As a response to this issue, we developed guidelines for manual sentence segmentation in the course of the LangBank project.[2] This project is dedicated to develop a digital infrastructure of Classical Latin and Historical German texts for scholars and learners. For this purpose, texts are digitalized and enhanced with various layers of manual, automatic, and semi-automatic linguistic annotations, most of which require a proper sentence segmentation. Therefore, these guidelines focus on a primarily syntactic approach with the purpose of facilitating further Natural Language Processing (NLP): The guidelines are designed to produce segments that are reasonable with respect to the contemporary research discussion on ENHG, yet also suited as input for NLP tools. Accordingly, the guidelines are influenced by linguistic as well as pragmatic considerations. As a result, our guidelines produce a variant of the classical t-units, which was modified to suit the special needs of ENHG. In the following, this variant is referred to as Early New High German t-unit (ENHG-TU). T-units were originally introduced by Hunt (1965, 20ff). They are usually defined as the 'shortest grammatically allowable sentences into which (writing can be split) or minimally **t**erminable **unit**'. They are well known in linguistic complexity and discourse analysis of both, spoken and written language (Lu 2010).

The guidelines were designed using the Register in Diachronic German Science (RIDGES) corpus as a reference. RIDGES is a tokenized multi-layer corpus of herbology texts ranging from the mid 15th to the late 19th century.[3] All following examples origin from the RIDGES corpus. However, the actual annotation guidelines as well as the in depth discussion were designed to be applicable to ENHG texts in general, which makes the guidelines applicable to various corpora.

In the following, first a list of all guidelines used to define ENHG-TUs is presented in section 2. Then, a detailed discussion of the separate rules is given in section 3, featuring example annotations and problematic cases.

[1] For a comprehensive overview over the obstacles of non-graphematic, philological sentence definitions, please see Schmidt (2016).

[2] `http://sfs.uni-tuebingen.de/langbank/index.html`.

[3] `http://korpling.german.hu-berlin.de/ridges/index_en.html`.

2 Overview

1. **Base definition:** An ENHG-TU consists of a phrasal head and all of its arguments and adjuncts and nothing else.
2. **Independence of the head:** The head of an ENHG-TU may not be the argument or the adjunct of another head itself, i.e. ENHG-TU do not govern each other.
3. **Uniqueness:** ENHG-TU may not overlap, i.e. no phrase is part of more than one ENHG-TU.
4. **Exhaustivity:** A text has to be partitioned exhaustively into ENHG-TUs.
5. **Minimal length I:** If
   a. the head of a potential ENHG-TU is structurally ambiguous with respect to its own status as argument or adjunct of another head, and
   b. it is not possible to disambiguate the structure based on textual coherence,

   the potential sentential unit is annotated as an ENHG-TU.
6. **Minimal length II:** If
   a. a phrase is structurally ambiguous with respect to its attachment to two ENHG-TUs, and
   b. it is not possible to disambiguate the structure based on textual coherence,

   the phrase is considered to be attached to the shorter ENHG-TU. The length of an ENHG-TU is defined in terms of tokens. If both ENHG-TUs in question contain the same amount of tokens, the phrase is attached to its preceding ENHG-TU.
7. **Finiteness:** An ENHG-TU includes preferably a finite verb. However, this is not mandatory. Therefore, if an ENHG-TU is ambiguous with respect to whether it contains a finite verb or not, the analysis including at least a single finite verb is to be preferred.
8. **Continuity:** ENHG-TUs are continuous strings of tokens. Discontinuous ENHG-TU are not possible, except if some meta text was inserted into a sentence.
9. **Sentence ending punctuation:** Unambiguously sentence ending punctuation has to be located at the outermost right periphery of an ENHG-TU. Unlike in contemporary German, whether punctuation is in fact sentence ending is highly dependent on register and period of origin of a given text.

3 Description

3.1 Base definition

Rule

An ENHG-TU consists of a phrasal head and all of its arguments and adjuncts and nothing else.

Description

This rule forms the shortest possible grammatical sentence, if possible: The finite verb of the main clause is the phrasal head and all arguments and adjuncts belonging to the verb's maximal projection are included. Arguments and adjuncts themselves may be phrases or subordinated clauses. Example 1 shows an ENHG-TU, whose head is a finite verb (*bedeutet*). It has a nominal phrase as its subject and a subordinate clause as its direct object.

(1) a. **Correct:** { auch **bedeutet** die rote Butte [...] / wie sehr zunahm die christliche Kirche auf dieser Erde von Christus gehalten lieb und wert }

b. **Wrong:** { auch **bedeutet** die rote Butte / [...] } { wie sehr zunahm die christliche Kirche auf dieser Erde von Christus gehalten lieb und wert } [4]

Because both phrases are in fact arguments of the finite verb, they have to belong to the same ENHG-TU. Coordinated main clauses, though, form separate ENHG-TUs. This is illustrated in example 2, which contains several coordinated main clauses, that have to be analysed as separate ENHG-TUs.

(2) a. **Correct:** { Galienus spricht dass das Kraut gut sei zu essen mit Lactuken, } { wann es sänftigt der Lactuken Kälte } { und sein Samen ist gut wider die Wassersucht, } { wann es erhitzt die Leber und reinigt sie ¶ }

b. **Wrong:** { Galienus spricht dass das Kraut gut sei zu essen mit Lactuken, wann es sänftigt der Lactuken Kälte und sein Samen ist gut wider die Wassersucht, wann es erhitzt die Leber und reinigt sie ¶ } [5]

Consistent with common linguistic analyses, the coordination is assumed to be the first token of the second conjunct in these cases. Note that *wann es erhitzt die Leber [...]* is a separate conjunct despite its adverbial function. As the last conjunct *und reinigt sie* ¶ does – unlike the other conjuncts – not contain a subject of its own, it does not qualify as a coordinated main clause, but as a coordinated verbal phrase. Hence, it is not annotated as a separate ENHG-TU: Only coordinated main clauses are separated into different ENHG-TU. Coordinated phrases or subordinate clauses remain in the same ENHG-TU as illustrated in example 3.

(3) a. **Correct:** { aber dieses Pulver soll er essen in der Speise: Braunwurz 6 Lot / Rheinblumen / Stechas citrina genannt / oder Krebsaugenstein 4 Lot / oder 4 Teile / Akeleiwurz 2 Lot / oder 2 Teile also fortgefahren. }

b. **Wrong:** { aber dieses Pulver soll er essen in der Speise: } { Braunwurz 6 Lot / } { Rheinblumen / Stechas citrina genannt / } { oder Krebsaugenstein 4 Lot / } { oder 4 Teile / Akeleiwurz 2 Lot / } { oder 2 Teile } { also fortgefahren. } [6]

This base definition mainly corresponds to the classical definition of t-units. However, it does not use the term "sentence", which is to be considered a benefit when it comes to defining sentence-like units.

[4] Rosbachs (1588, l. 3777ff)).
[5] Konrad von Megenberg (1482, l. 1566ff).
[6] Carrichter (1609, l. 3639ff).

It is important to note that the base definition does not require the phrasal head to be a finite verb, although for most cases it will be. However, in principle any phrasal head may be an ENHG-TU, if it is not governed by a finite verb. This is usually the case for headlines as in example 4, but also for all sorts of incomplete sentences as well as exclamations, interjections or appellations, as may be seen in example 5.[7]

(4) a. **Correct:** { Leibliche Nutzen und Wirkung. } { Dies Kräutlein Abbiss wird genannt [...] }
b. **Wrong:** { Leibliche Nutzen und Wirkung. Dies Kräutlein Abbiss wird genannt [...] } [8]

(5) a. **Correct:** { Ach und aber ach } { wie lang hat mich die Welt in die Finstere gezogen und lockt mich nach ihr }
b. **Wrong:** { Ach und aber ach wie lang hat mich die Welt in die Finstere gezogen und lockt mich nach ihr } [9]

It should also be noted, that in this respect the definition of ENHG-TU diverges from the original definition of t-units (Young 1995, p. 38), which cannot be interjections. However, since these types of phrases are neither semantically nor syntactically integrated to the actual clause, we decided to separate them. Constructions as the noun coordinations in example 3 do not constitute separate ENHG-TU, since they do not keep their coherent meaning in the text when separated from the phrase, see rule Minimal Length I for more details.

Trouble shooting

Coordination vs. subordination It is to be noted that coordinating connectives in contemporary German may be used as subordinating connectives in ENHG and vice versa. This is shown in example 6, where *doch* – a subordination in contemporary German– is used to coordinate two main clauses. Accordingly, it is analysed as the beginning of a new ENHG-TU. It also contains *denn*, which is a coordination in contemprary German, but used as a subordination in the example. Hence, the entire clause introduced by *denn* is part of the preceding ENHG-TU.

(6) { **doch** mögen wir Deutschen wohl und recht dafür gebrauchen das Kraut so man Welsamen nennt / **denn** es der Kraft / auch zum Teil der Gestalt nach dem rechten Seriphium ganz gleich ist } [10]

Shared arguments If two coordinated clauses share an argument, they are considered to be in the same ENHG-TU, if the structure can be analysed either as elision or as phrase level coordination.[11] This is illustrated in example 7.

(7) a. **Correct:** { Das Kraut entlöst Blähung in dem Leib und ~~das Kraut~~ öffnet das Verstopfen des Leibes } { und darum macht es schwitzen }
b. **Wrong:** { Das Kraut entlöst Blähung in dem Leib } { und öffnet das Verstopfen des Leibes } { und darum macht es schwitzen } [12]

The first two clauses are in the same ENHG-TU, because they share the subject *das Kraut*. The shared subject is indicated by repeating it right after the coordination and crossing it out: *das Kraut* is not

[7] Please see the subparagraph on appositions and parentheses in the trouble shooting paragraph of the rule of continuity in section 3.8 for further details.
[8] Rosbachs (1588, l. 50ff).
[9] Konrad von Megenberg (1482, l. 4905ff).
[10] Fuchs (1543, l. 304ff).
[11] Which of the two analyses is appropriate for a certain construction, is not relevant for our annotation and will, therefore, not be discussed. As elisions may be visualised by writing crossed-out covert phrases, the guidelines will sometimes assume elision for demonstrative purposes. However, this is motivated only by the concern that the common indication of phrase level coordination might be confusing, as it requires bracketing, which is already used to indicate ENHG-TU boundaries.
[12] Konrad von Megenberg (1482, l. 1506ff).

overtly realised at that position, but it is understood to be the subject even so. The following clause is in a separate ENHG-TU, because it contains its own subject: *es*. This also holds for all arguments, not only for subjects, as illustrated in 8.

(8) a. **Correct:** { Von dieser Wurzel getrunken räumt die Brust, ~~von dieser Wurzel getrunken~~ heilt die versehrte Lunge, ~~von dieser Wurzel getrunken~~ treibt aus den Kot }

b. **Wrong:** { Von dieser Wurzel getrunken räumt die Brust, } { heilt die versehrte Lunge, } { treibt aus den Kot } [13]

This rule also holds, if different grammatical functions are assigned to the shared argument, as illustrated in example 9, where *Blättlein* is object of the first and subject of the second finite verb.

(9) a. **Correct:** { das andere so Hühnerserb genannt wird / hat Blättlein ~~Blättlein~~ sind ein wenig rauher und ringsumher gekerbt [...] }

b. **Wrong:** { das andere so Hühnerserb genannt wird / hat Blättlein } { sind ein wenig rauher und ringsumher gekerbt / [...] } [14]

Still, these cases are analysed as a single ENHG-TU, as it is not clear, where the argument should belong, otherwise. Also, other approaches, for example assuming null subjects in these cases, cannot be reliably operationalized and leads to decreased quality of later NLP analyses.

Subordnate clause cluster In Early New High German subordinate clauses may occur in clusters without a governing main clause. In these cases all subordinate clauses are considered to form a single ENHG-TU. This is illustrated in example 10.

(10) a. **Correct:** { Da ich aber das Widerspiel erfahren / dass auf eine Zeit da der Samen unter dem / in die Reben getragenen Grund hervorgekommen / er über den Winter grün verblieben / und nachwärts im Sommer sehr groß geworden. }

b. **Wrong:** { Da ich aber das Widerspiel erfahren / } { dass auf eine Zeit da der Samen unter dem / } { in die Reben getragenen Grund hervorgekommen / } { er über den Winter grün verblieben / } { und nachwärts im Sommer sehr groß geworden. } [15]

[13] Adam von Bodenstein (1557, l. 1593ff).
[14] Fuchs (1543, l. 3395ff).
[15] Rhagor (1639a, l. 914ff)

3.2 Independence of the head

Rule

The head of an ENHG-TU may not be the argument or the adjunct of another head itself, i.e. ENHG-TU do not govern each other.

Description

It is necessary to impose this additional restriction, to prevent overgeneralisations based on the wide applicability of the base definition. Without further restrictions, all maximal projections of phrasal heads and subordinate clauses would qualify as ENHG-TUs. However, this is not intended. How the restriction applies to the segmentation process is illustrated in example 11.

(11) a. **Correct:** { auch **bedeutet** die rote Butte / darin der Samen liegt / wie sehr zunahm die christliche Kirche auf dieser Erde von Christus gehalten lieb und wert }

b. **Wrong:** { { auch } bedeutet die rote Butte / { darin der Samen liegt /} { wie sehr zunahm die christliche Kirche auf dieser Erde von Christus gehalten lieb und wert } } [16]

Troubleshooting

Currently, no issues regarding this rule are known.

[16] Rosbachs (1588, l. 3777ff).

3.3 Uniqueness

Rule

ENHG-TU may not overlap, i.e. no phrase is part of more than one ENHG-TU.

Description

This rule is consistent with the typical definitions of sentences and t-units. It mainly serves the purpose of giving a complete definition of ENHG-TUs. Example 12, which re-uses the sentence already introduced in example 9, illustrates the effect of this rule. To disambiguate the bracketing matching brackets were highlighted with colors in example 12b.

(12) a. **Correct:** { das andere so Hühnerserb genannt wird / hat Blättlein sind ein wenig rauher und ringsumher gekerbt [...] }

b. **Wrong:** { das andere so Hühnerserb genannt wird / hat { Blättlein } sind ein wenig rauher und ringsumher gekerbt / [...] } [17]

Troubleshooting

Currently, no issues regarding this rule are known.

[17] Fuchs (1543, l. 3395ff).

3.4 Exhaustivity

Rule

A text has to be partitioned exhaustively into ENHG-TUs.

Description

This rule is consistent with the typical definitions of sentences and t-units. It mainly serves the purpose of giving a complete definition of ENHG-TUs.

Troubleshooting

Currently, no issues regarding this rule are known.

3.5 Minimal Length I

Rule

If

a. the head of a potential ENHG-TU is structurally ambiguous with respect to its own status as argument or adjunct of another head, and

b. it is not possible to disambiguate the structure based on textual coherence,

the potential sentential unit is annotated as an ENHG-TU.

Description

This rule leads to more ENHG-TUs of shorter length. These are preferred for practical reasons, since NLP tools such as parsers work more efficient and less error prone with shorter sentential units. Also, ENHG is known to feature particularly long sentences, which exceed the average sentence length in contemporary German by far. Example 13 illustrates how the rule is applied.

(13) a. **Correct:** { aber von dem Baum des Erkenntnis Gutes und Böses sollst du nicht essen / } { denn welches Tages du davon isst / sollst du des Todes sterben }

b. **Wrong:** { aber von dem Baum des Erkenntnis Gutes und Böses sollst du nicht essen / denn welches Tages du davon isst / sollst du des Todes sterben } [18]

While one could argue for 13b due to the semantic relation between both potential ENHG-TUs, the *denn* clause is subordinated under the syntactically independent final clause *sollst du des Todes sterben*. Since 13a does not lead to an incoherent text and a lower maximum ENHG-TU length, it is the preferable analysis.

The restriction of the first rule of minimal length in 3.5 b. prevents overgeneralisations based on purely syntactic considerations. This is shown in example 14.

(14) a. **Correct:** { aber dieses Pulver soll er essen in der Speise: Braunwurz 6 Lot / Rheinblumen / Stechas citrina genannt / oder Krebsaugenstein 4 Lot / oder 4 Teile / Akeleiwurz 2 Lot / oder 2 Teile also fortgefahren. }

b. **Wrong:** { aber dieses Pulver soll er essen in der Speise: } { Braunwurz 6 Lot / Rheinblumen / Stechas citrina genannt / oder Krebsaugenstein 4 Lot / oder 4 Teile / Akeleiwurz 2 Lot / oder 2 Teile also fortgefahren. } [19]

Although the first part of the ENHG-TU from *aber* to *Speise* would already be saturated and the guidelines allow for solitary noun phrases as separate ENHG-TUs, it would clearly not be desirable to use the annotation in 14b.

Troubleshooting

Demonstrative vs. relative clauses It is to be noted that this rule leads to a systematic disambiguation between demonstrative and relative clauses, favouring the demonstrative clause analysis. These are ambiguous in ENHG and the disambiguation process is object of the contemporary research discussion (**???**). Example 16 shows an ambiguous case, where the clause in questions starts with a typical relative pronoun, combined with V2, which is actually a sign of demonstrative clauses. It should be noted at this point, that relative pronouns in ENHG may differ from contemporary German relativ pronouns. For

[18] Rosbachs (1588, l. 466ff).
[19] Carrichter (1609, l. 3639ff).

example *so* is a very common introduction for ENHG relative clauses, yet, in contemporary German, it cannot be used like that anymore. This is illustrated in example 15.

(15) { Unter denen erstlich M. Cato Censorius, von dem L.Columella meldet / dass er der erste gewesen / **so** den Feldbau die lateinische Sprache gelehrt }
(Rhagor 1639b, l. 169ff).

(16) a. **Correct:** { der gemeine Wermut ist ein Kraut mit vielen Zinken und Ästen / } { **an welchen** sind aschefarbene Blätter / vielfältig zerspalten / und goldgelbe Blumen / runder Same. }

b. **Wrong:** { der gemeine Wermut ist ein Kraut mit vielen Zinken und Ästen / **an welchen** sind aschefarbene Blätter / vielfältig zerspalten / und goldgelbe Blumen / runder Same. } [20]

Since demonstrative clauses are matrix clauses and relative clauses are subordinate clauses, the analysis of the given structure as a demonstrative clause is preferred due to the first rule of minimal length, even though, a primarily linguistically motivated distinction between these types of clauses might lead to other conclusions. Relative clauses are only annotated as such, when there are unambiguous relative clauses with finite verbs in the right sentence bracket. This is shown in example 17.

(17) a. **Correct:** { daher ich zu sagen pflege / dass man nicht nur auf den Lust / sondern viel mehr auf den Nutz pflanzen solle / als welches die rechte Endursache / und Zweck ist / darauf man sehen soll . }

b. **Wrong:** { daher ich zu sagen pflege / dass man nicht nur auf den Lust / sondern viel mehr auf den Nutz pflanzen solle / } { als welches die rechte Endursache / und Zweck ist / } { darauf man sehen soll . } [21]

To summarize, the first rule of minimal length allows for a systematic and consistent analysis regarding a highly problematic issue. However, this comes at the cost of not capturing the distribution of relative and demonstrative clauses in ENHG adequately. That is, research questions sensitive to that matter cannot be answered based on the information offered by ENHG-TUs, at least not without further adjustments in the search query used.

Independent vs. dependent clauses As stated in the previous subparagraph on demonstrative and relative clauses, the first rule of minimal length generally prefers matrix clause annotations over subordinate clause annotations in ambiguous cases. However, one criterion is used to identify subordinate clauses beyond any doubt, namely the position of the finite verb: if a finite verb is located in the right sentence bracket of a clause, it is considered to be a subordinate clause, i.e. it belongs to the ENHG-TU of its matrix clause. This finite verb does not necessarily have to be overtly realised, see afinite constructions discussed in rule 3.7. The position of the finite verb is considered to be a disambiguating factor: The verb position asymmetry between dependent and independent clauses was already established in Old High German (Abramowski 1979, p. 61, Axel 2007, 2009; Axel-Tober 2012, p. 284, Dittmer and Dittmer 1998). Hence, finite verbs in the right sentence bracket disambiguate the status of a clause, hence the first rule of minimal length does not apply in these cases. Example 18 shows such a case.

(18) a. **Correct:** { in griechischer und lateinischer Sprache wird es genannt Absinthium / welcher Name bis auf den heutigen Tag in den Apotheken geblieben **ist**. }

b. **Wrong:** { in griechischer und lateinischer Sprache wird es genannt Absinthium / } { welcher Name bis auf den heutigen Tag in den Apotheken geblieben **ist**. } [22]

[20] Fuchs (1543, l. 212ff).
[21] Rosbachs (1588, l. 621ff).
[22] Fuchs (1543, l. 40ff).

It shows a relative clause, which is introduced with *welcher Name*. Form the point of view of contemporary German, this is unexpected, because the proper relative pronoun referring to *Absinthium* would be *was*. *welcher Name* would be more indicating of a demonstrative clause. However, the position of the finite verb outranks these concerns.

When talking about the position of the finite verb, it is important to acknowledge the highly productive occupation of the postfield in ENHG. While in contemporary German, the postfield mainly contains subordinate clauses, in ENHG all elements from the middlefield could be postponed. This is shown in example 19.

(19) { [...] eine jede Blume ein Schote zwei / drei / herbringt / darin der Samen frei verborgen **liegt** eine lange Zeit } [23]

The term *verb-final*, which is also commonly used to describe the position of the finte verb in subordinate clauses, is therefore misleading. This also holds for the order of the verbs in the right sentence bracket: in contemporary German, the finite verb is – except for a few rare special cases – located at the rightmost periphery of the right sentence bracket, where several verbs might be clustered. However, in ENHG this order is less fixed. Hence, it is to be expected, that the finite verb is in fact located in the right sentence bracket, but not at its right periphery, as in example 20.

(20) { Apul. Und Orpheus lernen folgende Pillen zu machen / derer man alle morgens und abends zwei Skrupel dem Schwindsüchtigen **soll** geben / und gleich darauf **heißen** trinken laues Wasser. } [24]

Although neither constructions with occupied postfields nor constructions with a finite verb in the non-final position in the right sentence bracket look like verb last sentences, they in fact indicate subordinate clauses, because they are not positioned in the left sentence bracket.

Conditionals with *Spitzenstellung* The restriction that the first rule of minimal length only applies to cases, where the status of a clause is not disambiguated by text coherence, especially applies for conditional clauses. This is illustrated in example 21.

(21) a. **Correct:** { willst du diese Arznei kräftiger haben / nimm starken Weinessig zehn Lot / [...] }

b. **Wrong:** { willst du diese Arznei kräftiger haben / } { nimm starken Weinessig zehn Lot / [...] } [25]

This is a V1-conditional clause with a so called integrated *Spitzenstellung*, i.e. the protasis (*willst du diese Arznei kräftiger haben*) is located in the prefield of the aposis (*nimm starken Weinessig zehn Lot [...]*) (see König and Auwera 1988, p. 102). This construction is also well known in contemporary German. In this case, the protasis is dependent on the adoposis, as it is equivalent to the subordinate clause variant of a protasis with *wenn*. Therefore, all portases, whether the conditional clause is integrated or resumptive (adoposis with *dann*) are analysed as part of the ENHG-TU containing the adoposis.

This clearly differs from how other V1-clauses are treated, such as imperatives, which typically form separate ENHG-TUs:

(22) a. **Correct:** { willst du diese Arznei kräftiger haben / nimm starken Weinessig zehn Lot / ein halbes Lot der grünen Wermutblätter / Salz ein Drittel eines Quäntchens / } { mische diese Stücke wohl zusammen } { und trink es warm }

[23] Rosbachs (1588, l. 2060ff).
[24] Adam von Bodenstein (1557, l. 630ff).
[25] Mattioli (1563, l. 597ff).

b. **Wrong:** { willst du diese Arznei kräftiger haben / nimm starken Weinessig zehn Lot / ein halbes Lot der grünen Wermutblätter / Salz ein Drittel eines Quäntchens / mische diese Stücke wohl zusammen und trink es warm } [26]

[26] Mattioli (1563, l. 597ff).

3.6 Minimal Length II

Rule

If

a. a phrase is structurally ambiguous with respect to its attachment to two ENHG-TUs, and

b. it is not possible to disambiguate the structure based on textual coherence,

the phrase is considered to be attached to the shorter ENHG-TU. The length of an ENHG-TU is defined in terms of tokens. If both ENHG-TUs in question contain the same amount of tokens, the phrase is attached to its preceding ENHG-TU.

Description

This rule leads to more ENHG-TUs of shorter length. These are preferred for practical reasons, since NLP tools such as parsers work more efficient and less error prone with shorter sentential units. Also, ENHG is known to feature particularly long sentences, which exceed the average sentence length in contemporary German by far. Example 23 shows, how the rule is applied.

(23) a. **Correct:** { So bedeuten nun diese zwei Blätter / im Predigtamt die raue Lehre / das Gesetz in zwei Tafeln fest / } { von Gott gegeben auch zuletzt / die anderen drei grünen Blättlein / die drei Personen dir zeigen fein / im göttlichen Wesen }

b. **Wrong:** { So bedeuten nun diese zwei Blätter / im Predigtamt die raue Lehre / das Gesetz in zwei Tafeln fest / von Gott gegeben auch zuletzt / } { die anderen drei grünen Blättlein / die drei Personen dir zeigen fein / im göttlichen Wesen } [27]

According to all other rules, both segmentations are equally valid. However, for 23a the maximal ENHG-TU length is 20, while for 23b it is 26. According to the second rule of minimal length, this makes 23a preferable.

Troubleshooting

Currently, no issues regarding this rule are known.

[27] Rosbachs (1588, l. 3532ff).

3.7 Finiteness

Rule

An ENHG-TU includes preferably a finite verb. However, this is not mandatory. Therefore, if an ENHG-TU is ambiguous with respect to whether it contains a finite verb or not, the analysis including at least a single finite verb is to be preferred.

Description

This rule is intended to prefer ENHG-TUs with finite verbs over ENHG-TUs without finite verbs. This preference is necessary to state, because the rules of minimal length may collide with the assumption of shared arguments and it might not be clear which rule to follow. This is illustrated in example 24

(24) a. **Correct:** { Erdzwiebeln tun es auch Herzkraut ~~tut es~~ auch }
b. **Wrong:** { Erdzwiebeln tun es auch } { Herzkraut auch } [28]

24a is consistent with the assumption of shared arguments, 24b is consistent with the first rule of minimal length, textual coherence might be debatable. In this case, creating a single ENHG-TU instead of two ENHG-TUs where one is lacking a finite verb, is preferable. Given that the rule of minimal length was designed to facilitate parsing, this weighting of rules is reasonable, because 24b does not lead to successful parses either.

When applying this rule it is important to notice, that ENHG knows constructions featuring a covert finite verb and constructions that are actually without a finite verb. For example there are so called afinite constructions, where the finite auxiliary or – in rarer cases – modal verb does not have to be overtly realised, if it is positioned in the right sentence bracket in the topological field model. That is, afinite constructions are always VL clauses, cf. Reichmann and Ebert (1993, §256). This is illustrated in example 25. Also, elliptical copula constructions as in example 26 are especially common. In both examples, the covert finite verb was added but crossed out for illustrational purposes.

(25) { Unter denen erstlich M. Cato Censorius, von dem L.Columella meldet / dass er der erste gewesen ~~ist~~ / so den Feldbau die lateinische Sprache gelehrt ~~hat~~ }

(26) { Die Blättlein sind ein wenig rauher und ringsumher gekerbt / die Stängel ~~sind~~ purpurrot } [29]

Because in these cases the covert finite verb is highly salient, its overt absence qualitatively differs from actual constructions without finite verbs, such as headlines, interjections, etc., as seen in the base definition.

Troubleshooting

Currently, no issues regarding this rule are known.

[28] Carrichter (1609, l. 2014ff).
[29] Fuchs (1543, l. 3395ff).

3.8 Continuity

Rule

ENHG-TUs are continuous strings of tokens. Discontinuous ENHG-TU are not possible, except if some meta text was inserted into a sentence.

Description

This rule is primarily motivated by pragmatic considerations: most NLP tools work with continuous units. From a linguistic perspective, most sentences are in fact continuous, but they do not necessarily have to be so, as illustrated in example 27.

(27) Star Wars Episode vier – das habe ich schon immer gesagt – ist und bleibt der beste Star Wars Film.

Depending on the given sentence definition, one might argue that in fact two sentences are present in this example, because the parenthesis is an entire main clause on its own. However, based on this guidelines, the example shows only one ENHG-TU

Troubleshooting

Metatextual Insertions Meta texts such as glosses are not part of the text in a strict sense. Therefore, it is important to analyse the elements of different text layers independently of each other.

Appositions & Parentheses Unlike meta textual insertions, appositions and parentheses are part of the main text and, therefore, have to be analysed on the same level. Depending on the type of insertion, appositions and parantheses may qualify as ENHG-TU. The procedure is as follows:

If

1. the insertion does not qualify as ENHG-TU, it is analysed as a part of the surrounding or adjacent ENHG-TU.
2. the insertion does qualify as ENHG-TU , and if
   (a) the insertion is located at the left- or rightmost periphery of the ENHG-TU, it is to be analysed as a separate ENHG-TU.
   (b) the insertion is located within another ENHG-TU, it is to be analysed as part of this ENHG-TU.

Close appositions are, therefore, usually part of their surrounding ENHG-TU, while wide appositions and parantheses are analysed based on their position and internal structure. Example 28 shows an ENHG-TU-internal apposition, which actually qualifies as an ENHG-TU, as it may be considered as a complete grammatical sentence, except that it is surrounded by another ENHG-TU.

(28) a. **Correct:** { Wiewohl ich weiß dass nicht allein eben daran gelegen und genügsam ist / wann du dir Kräuter bringen lassest / und ihre Tugend in Büchern lesest / sondern dass auch / hoch notwendig zu wissen in welcher Zeit und an was Orte sie gewachsen sind / wann und wie ihre Einsammlung geschehen / in was Masse Proportion dieselbigen zum Brauch gegeben und dargereicht müssen werden / und welche in ihrem Leben / **das ist / wann sie noch grün und saftig sind** ihre Tugend bald erzeigen / ja auch welche nach ihrem Sterben / wann sie gedörrt dennoch von Gott Gnade empfangen kräftig ihre Eigenschaft zu erzeigen. }

b. **Wrong:** { Wiewohl ich weiß dass nicht allein eben daran gelegen und genügsam ist / wann du dir Kräuter bringen lassest / und ihre Tugend in Büchern lesest / sondern dass auch / hoch notwendig zu wissen in welcher Zeit und an was Orte sie gewachsen sind / wann und wie ihre Einsammlung geschehen / in was Masse Proportion dieselbigen zum Brauch gegeben und dargereicht müssen werden / und welche in ihrem Leben / { **das ist / wann sie noch grün und saftig sind** } ihre Tugend bald erzeigen / ja auch welche nach ihrem Sterben / wann sie gedörrt dennoch von Gott Gnade empfangen kräftig ihre Eigenschaft zu erzeigen. } [30]

Example 29 shows a similar case, where an address, usually considered to be a separate ENHG-TU is located inside of another ENHG-TU and, therefore, analysed as part of this ENHG-TU

(29) a. **Correct:** { nach langer Trübsal / **Christenleute** / in ewigem Leben/ auch bedeutet die rote Butte / darin der Samen liegt / wie sehr zunahm die christliche Kirche auf dieser Erde von Christus gehalten lieb und wert }

b. **Wrong:** { nach langer Trübsal / } { **Christenleute** / } { in ewigem Leben/ auch bedeutet die rote Butte / darin der Samen liegt / wie sehr zunahm die christliche Kirche auf dieser Erde von Christus gehalten lieb und wert } [31]

Another case is shown in 30. Here, the apposition is at the right periphery of an ENHG-TU, yet, it does not qualify as an ENHG-TU on its own.

(30) a. **Correct:** { hat Blätter gleich Quitten / allein, dass sie länger sind }

b. **Wrong:** { hat Blätter gleich Quitten / } { allein, dass sie länger sind } [32]

[30] Adam von Bodenstein (1557, l. 11ff)
[31] Rosbachs (1588, l. 3767ff).
[32] Adam von Bodenstein (1557, l. 366ff).

3.9 Sentence Ending Punctuation

Rule

Unambiguously sentence ending punctuation has to be located at the outermost right periphery of an ENHG-TU. Unlike in contemporary German, whether punctuation is in fact sentence ending is highly dependent on register and period of origin of a given text.

Description

This rule is the only acknowledgement to a graphematic sentence definition. Even with a consistent punctuation, which is not to be expected in ENHG, it is consistent with the classical definition of t-units to produce smaller units than graphematic sentences, for example in the case of coordinated matrix clauses. However, neither regular t-units nor ENHG-TUs are supposed to cross unambiguously sentence ending punctuation.

Also, this rule regulates the position of the punctuation between ENHG-TUs: they are located at the right periphery of the preceding ENHG-TU, not at the left periphery of the following ENHG-TU. Even if the punctuation initially is ambiguous between within and between ENHG-TU punctuation, after assigning ENHG-TU boundaries the punctuation becomes clearly sentence final and, therefore, belongs to the preceding ENHG-TU. This is illustrated in example 31.

(31) a. **Correct:** { Das heimisch Eppich ist wohlschmeckend / } { aber es ist dem Haupt böse und erweckt den wallenden Siechtum }

b. **Wrong:** { Das heimisch Eppich ist wohlschmeckend } { / aber es ist dem Haupt böse und erweckt den wallenden Siechtum } [33]

Troubleshooting

Unexpected Ambiguities Dots are not unambiguously sentence ending in ENHG, because they only emerged slowly in the 15th century. Thus, unlike in contemporary German, dots may end a sentence, but they can also occur within sentences, like commas. In return, commas may as well end sentences (Hartweg and Wegera 2005, p. 131). Example 32 shows such a non sentence ending dot.

(32) a. **Correct:** { das ist mehr für gelehrte Leute / als die sich mit dem Werk zu belustigen begehre. In welchem er aus Plinius sehr viel genommen }

b. **Wrong:** { das ist mehr für gelehrte Leute / als die sich mit dem Werk zu belustigen begehre. } { In welchem er aus Plinius sehr viel genommen } [34]

[33] Konrad von Megenberg (1482, l. 1525ff).
[34] Rhagor (1639c, l. 985ff).

Primary Literature

Adam von Bodenstein (1557). "Wie sich meniglich vor dem Cyperlin, Podagra genennet, waffnen solle unnd Bericht diser Kreüter, so den himmelischen Zeichen Zodiaci zugeachtet". In: Basel: Staehaelin. Chap. WieSichMeniglich_1557_vonBodenstein.xlsx, pp. 28–47.

Carrichter, Bartholomäus (1609). "Käutterbuch Des Edelen und Hochgelehrten Herren Doctoris Bartholomei Carrichters". In: Straßburg: Antonius Bertram. Chap. Kraeutterbuch_1609_Carrichter.xlsx, pp. 47–75.

Fuchs, Leonhart (1543). "New Kreüterbuch, in welchem nit allein die gantz histori, das ist namen, gestalt, statt vnd zeit der wachsung, natur, krafft vnd wuerckung, des meysten theyls der Kreüter so in Teuetschen vnnd andern Landen wachsen". In: Basel: Michael Isingrin. Chap. NewKreuterbuch_1543_Fuchs.xlsx, pp. 2–4.

Konrad von Megenberg (1482). "Das Buch der Natur". In: Augsburg. Chap. BuchDerNatur_1482_vonMegenberg.xlsx.

Mattioli, Pietro Andrea (1563). "New Kreüterbuch: Mit den allerschönsten vnd artlichsten Figuren aller Gewechsz, dergleichen vormals in keiner sprach nie an tag kommen". In: Prag. Chap. NewKreueterbuch_1563_Handsch.xlsx.

Rhagor, Daniel (1639a). "PflantzGart: Darinn grundtlicher Bericht zufinden, welcher gestalten 1. Obs-Gärten, 2. Kraut-Gärten, 3. Wein-Gärten". In: Bern: Ben Stephan Schmid. Chap. PflantzGart-c4_1639_fixed.xlsx, pp. 33–45.

Rhagor, Daniel (1639b). "PflantzGart: Darinn grundtlicher Bericht zufinden, welcher gestalten 1. Obs-Gärten, 2. Kraut-Gärten, 3. Wein-Gärten". In: Bern: Ben Stephan Schmid. Chap. PflantzGartVorrede_1639_Rhagor.xl pp. 1–10.

Rhagor, Daniel (1639c). "PflantzGart: Darinn grundtlicher Bericht zufinden, welcher gestalten 1. Obs-Gärten, 2. Kraut-Gärten, 3. Wein-Gärten". In: Bern: Ben Stephan Schmid. Chap. PflantzGart _1639_Rhagor.xlsx, pp. 92–110.

Rosbachs, Konrad (1588). "Pardadeißgärtlein Darinnen die edleste unnd fürnehmbste nach ihrer Gestalt und Eigenschaft abcontrafeytet". In: Frankfurt am Main: Johann Spieß. Chap. Paradeiszgaertlein_1588_Rosbach.xlsx, pp. 1–43.

Secundary Literature

Abramowski, Anneliese (1979). "Der Beitrag der Beschwerdeschriften aus der Zeit des Bauernkrieges in Deutschland 1525/26 zur Herausbildung einer nationalen Norm der Literatursprache unter besonderer Berücksichtigung syntaktischer Entwicklungstendenzen". PhD thesis.

Axel, Katrin (2007). *Studies on Old High German Syntax. Left Sentence Periphery, Verb Placement and Verb-Second*. Amsterdam, Philadelphia: Benjamins.

Axel, Katrin (2009). "The verb-second property in Old High German: Different ways of filling the prefield." In: *New Approaches to Word Order Variation and Change in the Germanic Languages*. Berlin, New York: de Gruyter, pp. 17–44.

Axel-Tober, Katrin (2012). *(Nicht-) kanonische Nebensätze im Deutschen: synchrone und diachrone Aspekte*. Vol. 542. Walter de Gruyter.

Dittmer, A. and E. Dittmer (1998). *Studien zur Wortstellung – Satzgliedstellung in der althochdeutschen Tatianübersetzung. Für den Druck bearbeitet von Michael Flöer und Juliane Klempt*. Göttingen: Vandenhoeck & Ruprecht.

Hartweg, Frédéric and Klaus-Peter Wegera (2005). "Frühneuhochdeutsch: eine Einführung in die deutsche Sprache des Spätmittelalters und der frühen Neuzeit". In: *Germanistische Arbeitshefte*. 2. Auflage. Vol. 33. Walter de Gruyter.

Hunt, Kellogg W. (1965). “Grammatical Structures Written at Three Grade Levels”. In: *NCTE Research Report* 3.

König, Ekkehard and Johan Van der Auwera (1988). “Clause integration in German and Dutch conditionals, concessive conditionals and concessives”. In: *Clause combining in grammar and discourse*. Ed. by John Haiman and Sandra A. Thompson. Vol. 18. John Benjamins Publishing, pp. 101–133.

Lu, X. (2010). “Automatic analysis of syntactic complexity in second language writing”. In: *International Journal of Computational Linguistics* 15.4, pp. 474–496.

Reichmann, Oskar and Robert Peter Ebert (1993). *Frühneuhochdeutsche Grammatik*. Vol. 12. Sammlung kurzer Grammatiken germanischer Dialekte. A, Hauptreihe. Tübingen: Niemeyer.

Schmidt, Karsten (2016). “Der graphematische Satz”. In: *Zeitschrift für germanistische Linguistik*, pp. 215–256.

Young, Richard (1995). “Conversational Styles in Language Proficiency Interviews”. In: *Language Learning* 45.1, pp. 3–42.