JESLA | Journal of the European Second Language Association

Mylläri, T. (2020). Words, clauses, sentences, and T-units in learner language: Precise and objective units of measure? *Journal of the European Second Language Association,* 4(1), 13–23. DOI: https://doi.org/10.22599/jesla.63

RESEARCH

Words, clauses, sentences, and T-units in learner language: Precise and objective units of measure?

Taina Mylläri

In research on learner language complexity, accuracy and fluency (CAF), syntactic complexity is often studied with quantitative measures based on words, clauses, sentences, and T-units. The findings have been mixed, but segmenting learner language into these units of measure has seldom been problematised, even if the need for accurate coding is well known. The present study explores words, clauses, sentences, and T-units as production units in written learner language using a corpus of 352 L2 Finnish texts (28,813 words). The results illustrate how written learner language can be hard to fit into the production unit categories, which are essential for the most frequently used quantitative measures of syntactic complexity. On the one hand, the results support calls to include explicit definitions of the units of measure when reporting findings obtained with these quantitative measures. On the other hand, they align with calls to introduce new measures to better gauge the changes in learner language syntax as it develops with increasing language proficiency.

Keywords: Common reference levels; Complexity; Learner Finnish; Learner writing; Segmentation

1. Introduction

When second-language (L2) learning is analysed in terms of complexity, accuracy, and fluency, complexity is often quantified using measures that are based on the length of clauses, sentences, and T-units, or on the relation of these production units to each other (e.g., Bulté & Housen, 2012; Pallotti, 2015; Wolfe-Quintero et al., 1998). These measures require the consistent and reliable segmenting of learner language, but the possible effects of inconsistencies in coding learner language have seldom been discussed (e.g., Byrnes et al., 2010, p. 169).

Learner language does not always fit neatly into the categories used in these quantitative measures of complexity. Deviations from the target language norms are a challenge for annotation (e.g., Granger, 2002), and there can be several interpretations of the intended target form (e.g., Brunni et al., 2015; Ragheb & Dickinson, 2011; Rehbein et al., 2012). These challenges affect the segmenting of learner language into clauses, sentences, and T-units, especially on lower proficiency levels, when learner language can be fragmented and elliptic in both its oral (e.g., Foster et al., 2000) and written forms (e.g., Martin, 2013). The ambiguity of clause and sentence boundaries in written learner language is illustrated by Martin's (2013) segmenting experiment, in which a group of 35 university students of Finnish segmented three learner Finnish texts into clauses and sentences. The results showed variation in the numbers of both sentences and clauses, and even when two students arrived at the same number of clauses or sentences, the production units identified were not necessarily identical (Martin, 2013).

Differences in the numbers of production units are likely to lead to different results when complexity is measured using these units. Segmenting learner language into clauses, sentences, and T-units may also affect the quantitative measures that have typically been used to measure the syntactic complexity of written learner language, as among the most frequently used measures have been mean length of sentence, mean length of clause, mean length of T-unit, mean number of clauses per T-unit, mean number of T-units per sentence, and mean number of dependent clauses per clause (e.g., Ortega, 2003).

The present study seeks to explore how objective and reliable words, sentences, clauses, and T-units are as units of measure in written learner language. This is done by taking a close look at the segments that cause difficulties in splitting the data into these production units. The research question is: How do deviations from target language norms affect the segmenting of written learner language into words, sentences, clauses, and T-units? To answer this question, a corpus of written learner Finnish texts from different proficiency levels, from beginners to advanced, was segmented into these production units, and the segments not fitting into these categories were analysed. While the results are in part language specific, the problems are not limited to learner Finnish: Similar problems arise with other languages too.

Department of Language and Communication Studies, University of Jyväskylä, FI

taina.myllari@jyu.fi

2. Word, sentence, clause, and T-unit as production units

When words, clauses, sentences, and T-units are used as units of measure, they need to be identified in the data and their frequency of occurrence needs to be counted. These units can, however, be defined in more ways than one. In this section, words, clauses, sentences, and T-units are discussed in relation to their use in measuring syntactic complexity.

2.1. Word

One way to measure complexity is to calculate the mean length of a given production unit in words (e.g., Bulté & Housen, 2012). In many languages, a word can be defined as an orthographic unit separated from other text units by a blank space or by punctuation. While this simple definition is not suitable for all languages and it may overlook some linguistic features of words and differences between languages (e.g., Booij, 2012), it can in many cases be considered a reasonable way of defining a word in written language (Haspelmath, 2011, p. 69). It also makes automated word counts easy in languages in which words are separated by blank spaces.

This simple definition of a word seems reasonable within a study or within a language, but some language-specific conventions or orthographic rules, such as those concerning compound words, may cause differences in word count. When the number of words is based on orthography, elements in compound words are each counted as one word if they are separated from other elements by a blank space. This way of counting seems suitable for the present study, as compound words in Finnish normally consist of two or more words spelled as one orthographic unit (e.g., *ruokapöytä* for *ruoka+pöytä* 'food' + 'table') 'dining/dinner table'. It may, however, cause problems in languages with different orthographic conventions. Additionally, errors in orthography with compound words made by both L2 and first-language (L1) writers, such as *iso äiti* for *isoäiti* 'grandmother' or *jokapäivä* for *joka päivä* 'every day', may affect the word count.

Another possible source of differences in the length of a clause, sentence, or T-unit in words are differences in morphology. In morphologically rich languages, some syntactic information may be encoded within a single word, as illustrated in example (1). Such differences, and their impact on word count, should be taken into consideration if the length of a given syntactic unit in words is compared across languages.

(1)	talo-ssa=ni	luk-isi-t=ko
	house-INESS=POSS.1SG	read-COND.2SG=Q
	'in my house'	'would you read'

Some less-frequently occurring elements in written texts may also affect the word count. These include abbreviations pointing to multiple words (e.g, *jne* for *ja niin edelleen* 'and so on'), orthographic units containing hyphens or slashes, and word-like units containing or consisting of other characters than letters of the alphabet, such as expressions of quantity written with numbers (e.g., *1–2*), or amounts specified with a combination of a number and a unit of measurement (e.g., *12 tuntia* '12 hours'; *11 tuntia* '11 hours').[1]

2.2. Clause

Some of the most widely used measures of syntactic complexity involve counting the number of clauses per given unit (Pallotti, 2015) and mean length of clause in words (e.g., Ortega, 2003). Although grammars offer relatively clear definitions of a clause, in reality texts, both in L1 and L2, contain segments that do not fit these descriptions. Nevertheless, these segments should also somehow be acknowledged and included in analyses of complexity.

In studies on syntactic complexity in learner language, especially in learner English, a clause has typically been defined as a production unit containing either a subject and a finite verb or a subject and a finite or non-finite verb form (e.g., Lu, 2011, p. 44; Wolfe-Quintero et al., 1998, p. 70). When measuring syntactic complexity, infinitive forms in verb clusters can be considered to either belong to a verb construction within one clause or to form non-finite dependent clauses (e.g., Pallotti, 2015). In Finnish, structures with a non-finite verb form are typically considered verb phrases rather than clauses (Hakulinen et al., 2004, pp. 488–489; Vilkuna, 2003, pp. 14–15). Regarding the measures of complexity, coding verb clusters to belong to one clause or to more clauses has an impact on the mean length of clause, as well as on the number of clauses (Bulté & Housen, 2012). This decision also affects the number of dependent clauses and thus any ratios in which the number of dependent clauses is used.

In the above definitions of a clause, a subject is also considered a mandatory element. While this requirement suits non-null-subject languages, such as English, it is not practical for null-subject languages or partial null-subject languages, such as Finnish. In a quantitative study of Finnish syntax, Hakulinen et al. (1996) conclude that an overt subject cannot be considered a mandatory element of a clause in Finnish, because in their data, consisting of factual prose such as newspaper articles, more than 30% of the clauses did not have an overt subject (Hakulinen & Karlsson, 1980). There are several linguistic features contributing to this. In Finnish, it is possible to incorporate the first- and second-person subject in the verb form, leaving out the corresponding pronoun. Hence, for example, 'I say' can be expressed either with two words (*minä sanon*) or one word (*sanon*). There are also clause types that do not allow an overt subject. These types include all clauses in the passive voice (Hakulinen et al., 2004, p. 1245; Karlsson, 2015, p. 200) and some clauses containing meteorological expressions (e.g., *Satoi.* rain-PAST-3SG 'It was raining.') or causative verbs (e.g., *Minua pelottaa.* me-OBJ frighten-PRS-3SG 'I feel frightened.') (e.g., Karlsson, 2015, p. 81; for more detail, see Hakulinen et al., 2004, pp. 856–862, 1286). Such differences between languages need to be considered when defining a clause.

2.3. Sentence

In segmenting written language, the sentence can be considered "the obvious unit" (Ellis & Barkhuizen, 2005, p. 147). A sentence is usually defined as an orthographic unit beginning with a capital letter and ending with appropriate punctuation. These indicators of sentence boundaries are marked by the writer, but in some texts, the use of punctuation and capital letters may be inconsistent. These inconsistencies may be caused by problems in writing in the target language or by problems in writing in general.

The unsystematic use of punctuation can sometimes create sentences without a verb (as in example (2)) or an apparent independent clause (see example (3)). Considering this kind of punctuation intentional or erroneous affects the number of sentences and the kind of elements they consist of.

(2) Saa syödä purukumia tunnilla ellei se
can eat chewing.gum in.class unless.not it
häiritse. muita.
disturbs others.
'You/One can eat chewing gum in class unless it disturbs. others.' (F-010, adolescent A1)

(3) Oppilaat eivät sais ottaa kännyköitä kouluun
pupils not should take mobiles to.school
mukaan. Koska ne häiritse tunneilla.
along because they disturb in.classes
'Pupils should not take mobile phones to school. Because they disturb the class.' (F-733, adolescent B1)

Not all sentences without a verb or an independent clause result from errors in punctuation. For example, newspaper headlines, interactive elements such as greetings, and certain idiomatic expressions can be punctuated as sentences even when they do not contain a grammatically complete clause (e.g., Biber et al., 1999, pp. 224–225; Leech & Svartvik, 2002, p. 262). This also applies to Finnish. According to standard Finnish grammar, the minimal length of a sentence is one word, and this word does not need to be a verb (Hakulinen et al., 2004, p. 827).

There are also sentences that contain only clauses or structures that are traditionally not considered independent. For example, Foster et al. (2000) raise the question of the dependence or independence of adverbial clauses beginning with the conjunction *because* but lacking an apparent main clause. In written Finnish, sentences containing only clauses that begin with a subordinator can be found in both L1 and L2 writers' texts (Kalliokoski, 2006). In Finnish, there are also sentences that contain only infinitive verb forms (Visapää, 2008).

Sentences containing grammatically incomplete clauses or lacking an independent clause present a challenge to coding learner language and to the quantitative measures of complexity. Annotating these sentences to contain at least one clause or zero clauses affects all measures in which the number of clauses is used. Similarly, coding these sentences to contain at least one independent clause or only dependent clauses also affects measures relying on the number of dependent or independent clauses.

2.4. T-unit

The T-unit, first introduced by Hunt in 1965 in the L1 context, has gained ground in L2 research, but it has also been the target of some criticism (Bardovi-Harlig, 1992; Biber et al., 2011; Crossley & McNamara, 2014). There are several definitions of the T-unit. Most often it refers to one independent clause and any dependent clauses attached to it, although there has been variation in the inclusion or exclusion of fragments and in the counting of elements across sentence boundaries (e.g., Foster et al., 2000, pp. 360–363). In measuring syntactic complexity, the T-unit is among the most popular production units (Foster et al., 2000; Ortega, 2003; Wolfe-Quintero et al., 1998).

However, the relationship between clauses can sometimes be ambiguous, which makes it hard to determine whether a clause is coordinated or subordinated (Lieko, 1992, pp. 29–31; Quirk et al., 1972, pp. 795–796). Additionally, it is not always clear which independent clause is the main clause of a given dependent clause (as in example (4)), where it is not clear which of the independent clauses functions as the main clause for the clause beginning with *jos* 'if'.

(4) jos kotona on kiire, valmistan ruokaa, ja
if at.home is hurry I.make food and
huomasin että ei ole maitoa, menen
I.noticed that no is milk I.go
lähikauppaan.
to.corner.shop
'if it's busy at home, I cook, and I noticed that there is no milk, I go to the corner shop.' (F-253, adult A2)

Nevertheless, distinguishing between the two and identifying the dependency relationships are essential when using the T-unit as a unit of measure.

3. Design of the study

In the present study, a corpus of written learner Finnish and a comparative set of L1 Finnish adolescent writers' texts were split into words, sentences, clauses, and T-units to create a corpus for measuring syntactic complexity in learner Finnish with the frequently used quantitative measures. To find the production units, a set of definitions, described in Section 4, was used, and segments not fitting into these categories were examined. The focus was on problematic segments that could lead to different interpretations of the number of the relevant production units (i.e., words, sentences, independent clauses, and dependent clauses). The problematic segments were analysed qualitatively and quantified by counting their frequency. The aim was to identify the key challenges and evaluate their significance.

3.1. The data

The data in the present study comprise 352 learner Finnish (L2) texts (28,813 words) and 128 native Finnish (L1) texts (7,049 words) from the Cefling project corpus,[2] which contains texts elicited by means of communicative writing tasks. The Cefling corpus was collected for L2 research by selecting L2 Finnish adult learner texts from the National Certificates of Language Proficiency examination database and by collecting texts from adolescent L2 Finnish learners and L1 writers in school years 7 to 9 (age 12 to 16) with matching tasks (Martin et al., 2010). For the present study, the argumentative texts from the Cefling corpus were used.

To facilitate research into the development of different linguistic features in relation to language proficiency, all the L2 Finnish texts were assessed and placed according to the proficiency levels of the Common European Framework of Reference (CEFR, Council of Europe, 2001) by a team of trained raters in the Cefling project. Each text was rated by three raters using scales based on the CEFR (Alanen et al., 2010). The reliability of the ratings has been shown by both quantitative and qualitative analysis (for more detail, see Huhta et al., 2014). The adult learners' texts cover CEFR proficiency levels A1 to C2, and the adolescent learners' argumentative texts cover levels A1 to B1.

In the present study, segments that were copied word by word from the task prompts or contained only verbless greetings, pseudonyms, or contact information were considered echo responses and interactional elements, and they were not included in the analysis (cf. Foster et al., 2000). This led to the exclusion of 328 segments (961 words). The remaining text in the Cefling project Microsoft Word files was organised into a project corpus (**Table 1**).

To enable comparisons between language learners and native speakers, the L2 and L1 data were kept separate. To observe differences between learner age groups and between proficiency levels, the L2 data were separated into two groups, referred to in this study as adult learners and adolescent learners, and arranged according to the assessed proficiency level. Similarly, the L1 data were organised into three subgroups based on the school year of the participants.[3]

3.2. Analysis of the data

To answer the research question, the data were coded as words, sentences, clauses, and T-units. Segments not complying with the definitions and thus not fitting into these categories were analysed linguistically, and the frequency of such segments was calculated. On the sentence level, the focus was on irregularities in sentence marking which could affect the number of clauses, sentences, and T-units. On the clause level, the focus was on segments that could affect the number of clauses or their status as independent or dependent. If the problematic segments were not considered to affect the number of production units or the division of clauses into independent and dependent, they were outside the scope of this study.

Because there was only one annotator and a high number of problematic segments were found during coding, the sentence-level segmentation was compared with two other segmentations of the same data. The segmentation in the Cefling project CHAT files was one of those used. During the Cefling project, the texts were divided into sentences by seven native Finnish-speaking graduate students pursuing their Master's degree in Finnish language. If a sentence could not be clearly identified, the students were instructed to divide the text into clauses or, if the clause boundaries were also ambiguous, to group the text into clauses around the finite verbs (Cefling project, unpublished instructions). In the Cefling project, problematic segments were discussed but no inter-annotator agreement was counted or reported. The second segmentation used the open-source dependency parsing pipeline for Finnish developed by the University of Turku natural language processing (NLP) group.[4] The Finnish Dependency Parser is a statistical parser based on open-source NLP tools and trained on the Turku Dependency Treebank, whose system of annotation is a

Table 1: Amount and distribution of data across different writer groups.

CEFR level/ school year	Adult		Adolescent		Native		Total	
	texts	words	texts	words	texts	words	texts	words
A1	50	2,261	32	775	–	–	82	3,036
A2	37	2,272	39	1,589	–	–	76	3,861
B1	43	5,142	40	2,232	–	–	83	7,374
B2	35	4,166	–	–	–	–	35	4,166
C1	46	5,876	–	–	–	–	46	5,876
C2	30	3,879	–	–	–	–	30	3,879
Year 7	–	–	–	–	55	2,902	55	2,902
Year 8	–	–	–	–	50	2,831	50	2,831
Year 9	–	–	–	–	23	976	23	976
Total	241	23,596	111	4,596	128	6,709	480	34,901

Finnish-specific adaptation of the Stanford Dependency scheme (Haverinen et al., 2014).

To evaluate the reliability of the sentence-level segmentation, the three segmentations were compared using precision, recall, and F-score, which is the harmonic mean of the two. None of the segmentations was used as a gold standard annotation but instead, precision and recall were counted following Lu (2010) and Brants (2000) by dividing the number of segments identical in both the compared sets by the total number of sentences in the first set (precision) and in the second set (recall). In this kind of comparison setup, precision, recall, and F-score are considered to reflect agreement between annotations, the F-score being considered the most informative of the three (Brants, 2000; Lu, 2010).

4. Results

4.1. Words

In the present study, a word was defined as an orthographic unit containing alpha-numeric characters and separated from other units by a blank space, punctuation, or other orthographic marker, such as the beginning or the end of a line or a paragraph.

During the sentence-level comparisons, the orthography of each word in the two manual segmentations was checked and aligned to eliminate inconsistencies due to typing errors or differences in typing conventions between the file formats. Any discrepancies were resolved, when possible, based on the hand-written originals (adolescent learners and L1 writers) or the original database files (adult learners), and otherwise based on the transcription in the Word files. This resulted in identical word counts in the two manual segmentations.

In the automatically segmented data, there were four words more in the adult learner data and two words more in the L1 data than in the manual segmentations. The differences were caused by non-alphabetic characters within a word, such as quotation marks or a colon connecting a letter and a case ending. There were no differences in the word count in the adolescent learner data.

4.2. Sentences

A sentence was initially defined as an orthographic unit beginning with a capital letter and ending with a full stop, question mark, exclamation mark, or any combination of these. However, the requirement of initial capitalisation was discarded during segmenting because in some texts all the writing was originally in block capitals, or random block capitals were used within words. Consequently, segments such as those in example (2) were also coded to contain two sentences. The requirement of punctuation at the end of a sentence was also re-evaluated, and other orthographic markers, for example the organisation of text into items on a bulleted or numbered list, were considered to be indicators of sentence boundaries, as some texts were partly or completely organised as lists (as in example (5)).

(5) Minä olin syömässä ravintolassa Helsingissa, minä nähnyt 3 huonoa asiaa ja 1 hyvä asia
1/- ruokaa on hyvää.
2/- paljon ihmiset, ei riita paikkalla,
3/- He puhuvat kovaa
4/- ravintolassa tosi kuuma.
'I was eating at a restaurant in Helsinki, I seen 3 bad things and 1 good thing
1/- food is good.
2/- a lot of people, no quarrel at place,
3/- They speak loudly
4/- at the restaurant really hot.' (F-1012, adult A1)

In example (5), which is a short text from the lowest proficiency level, there is only one sentence indicated with both initial capitalisation and punctuation at the end. After careful consideration of such cases, the working definition of a sentence was changed, and the end of a whole text, a text paragraph, or a list item in a bulleted or numbered list were also defined as ending a sentence, regardless of the punctuation.

To evaluate the effect of the changes in the definition of a sentence, the sentence-level segmentation was compared to the original definition, and sentences not falling within the original definition were divided into two categories: Those ending with standard punctuation but not beginning with a capital letter, and those having no standard punctuation at the end (**Table 2**). The comparison showed that with proficiency level A1, only around half of the sentences conformed to the original definition of a sentence. Inconsistencies in punctuation were more frequent in the learner texts than in the L1 texts, where they were rare. These results should not, however, be interpreted as a straightforward relationship between the use of punctuation and L2 proficiency, as the inconsistent use of punctuation may have been caused by difficulties in writing in general, not necessarily difficulties in writing in a L2.

As for the actual number of sentences, there were only small differences in the numbers found in the different segmentations, and agreement between the segmentations was high, 90% to 99%, except in the adolescent learner data, where it was 85% and 88% on levels A1 and A2 in the comparison of the two manual segmentations (**Table 3**). The high agreement indicates that the sentences found were mainly identical.

The Cefling project segmentation contains the highest number of sentences in all the writer groups, which is in line with the instructions to split the text into clauses if the sentence boundaries were unclear. The parsed texts were found to contain the smallest number of sentences in all the writer groups. According to Haverinen et al. (2014), the parser makes its decisions based on dependencies and does not follow any separately given rules for sentence splitting.

These results seem to suggest that the working definition used in the present study could provide reliable enough criteria for identifying a sentence. It seems that

Table 2: Number and percentage of sentences with initial capitalisation and standard punctuation, sentences ending with standard punctuation but lacking initial capitalisation, and sentences not ending with standard punctuation.

Writer group	Initial capital and standard punctuation		Standard punctuation but no initial capital		No standard punctuation		Total sentences
	n	%	n	%	n	%	n
Adult A1	154	48	60	19	108	34	322
Adult A2	234	63	29	8	109	29	372
Adult B1	414	88	30	6	25	5	469
Adult B2	378	95	6	2	12	3	396
Adult C1	537	98	4	1	9	2	550
Adult C2	367	96	3	1	12	3	382
Adolescent A1	48	55	15	17	24	28	87
Adolescent A2	118	86	13	9	6	4	137
Adolescent B1	182	93	8	4	5	3	195
L1 year 7	268	93	8	3	11	4	287
L1 year 8	254	92	10	4	12	4	276
L1 year 9	104	94	2	2	5	5	111

Table 3: Number of sentences in each segmentation and the results of the sentence-level comparisons.

Writer group	Number of sentences in different segmentations			Number of identical sentences		Agreement between segmentations (F-score)	
	Present study	CHAT files	Parsed texts	Present study and CHAT files	Present study and parsed texts	Present study vs. CHAT files	Present study vs. parsed texts
Adult A1	322	337	308	308	296	0.93	0.94
Adult A2	372	375	371	364	368	0.97	0.99
Adult B1	469	488	450	452	430	0.94	0.94
Adult B2	396	405	385	387	371	0.97	0.95
Adult C1	550	554	545	546	530	0.99	0.97
Adult C2	382	388	379	377	367	0.98	0.96
Adolescent A1	87	97	84	78	81	0.85	0.95
Adolescent A2	137	145	131	124	123	0.88	0.92
Adolescent B1	195	201	192	185	183	0.93	0.95
L1 year 7	287	290	280	285	259	0.99	0.91
L1 year 8	276	277	268	273	244	0.99	0.90
L1 year 9	111	112	106	110	101	0.99	0.93

the absence of an initial capital letter can be ignored. Further, the end of a list item in a bulleted or numbered list, the end of a text paragraph and the end of the whole text could be considered indicators of a sentence ending, even if none of these markers are included in the standard definition of a sentence.

4.3. Clauses

A clause was defined as a segment within a sentence containing a finite verb and all its arguments and adjuncts. As Finnish is considered a partial null-subject language, a subject was not required. Following the definition in Hakulinen et al. (2004, pp. 827–828), a finite verb was deemed to be a mandatory element in a clause, and non-finite verbs were considered to be part of a verb phrase within a clause clustered around a finite verb, although in some studies (e.g., Hakulinen et al., 1996) or descriptions of Finnish grammar (e.g., Karlsson, 2015) also some structures clustered around non-finite verb forms have been considered clauses. As the texts were first split into sentences, and this segmenting was considered reasonably reliable, it was decided to look for clauses within sentences.

However, splitting the data into clauses proved to be problematic. In the first place, not all sentences contained a grammatical clause. In some sentences, especially with the lower proficiency levels, verbs could be completely missing or determining the presence or absence of finite verbs could require interpretation. Some of these verbless sentences were created by punctuation that seemed to split a grammatical clause into two sentences (as in example (2)). Others, especially among the higher proficiency levels, seemed to be stylistically motivated and to intentionally lack a finite verb (see example (6)). With some of these sentences, context was needed in order to choose between several interpretations (as in example (7)), in which the words *soitin* (musical_instrument.NOM or call.PAST.1SG) and *vasta* 'just' could have more than one meaning and could be labelled as more than one part of speech: The word *vasta* could also be a misspelled form of *vasta-a* (answer.PRS.3SG or answer.INF). Additionally, there were sentences containing only non-finite verb forms, such as infinitives (example (8)), participles, or a negation verb without the main verb.[5]

(6) Ensimmäinen työpäivä ja hetkessä se onkin ohi. Sitten viikko ja kuukausi.
'First day of work and suddenly it's over already. Then a week and a month.' (F-816, adult C1)

(7) Sitten sinä vasta puhelin
then you just/answer phone
soitin.
musical.instrument/I.called
'Then you answer the ringing phone.' (A possible interpretation) (F-249, adolescent A1)

(8) Kävel-lä luontossa, katso-a kauniita paikkoja,
walk-INF in.nature look-INF beautiful places
nautti-a meren- tai järven vettä.
enjoy-INF of.sea- or of.lake water
'To walk in nature, look at beautiful places, enjoy the sea or lake water.' (F-659, adult B1)

It was also problematic because in sentences with more than one finite verb, it was not always clear how many clauses the finite verbs should be divided into. As in example (9), there could be two finite verbs (i.e., *ei saa* 'may not' and *saavat* 'may'), but it was not clear if there were two clauses.

(9) ei saa lapset saa-vat ol-la kauan
not get[PRS.3SG] children get-PRS.3PL be-INF long
nettissä
on.the.web
'may not children may be on the internet for a long time.' (F-018, adolescent A1)

Thirdly, coordinators and subordinate conjunctions were sometimes used to connect segments that did not fall within the definition of a clause. As coordinators can be used to connect both clauses and phrases, segments without a finite verb could be interpreted as phrases coordinated with an element in the preceding clause. Another interpretation could be, as in example (10), that there are two coordinated clauses of which the latter is elliptic: The word *kielettyä* 'forbidden' could be interpreted as an adjective coordinated with *sallittua* 'allowed' in the preceding clause or as an elliptic clause *mutta [että kännykän pitely on] koulussa kielettyä* 'but [that holding a mobile is] at school forbidden'.

(10) toivomme että, kännykän pitely on sallittua,
we.hope that a.mobile holding is allowed
mutta koulussa kielettyä.
but at.school forbidden
'we hope that, holding a mobile is allowed, but at school forbidden.' (F-736, adolescent A2)

Regarding the use of subordinate conjunctions, this could create dependent clauses without a grammatical main clause (as in example (11)) or elements beginning with a subordinator but not containing a verb (see example (12)). We will return to this issue when exploring the T-units in the data.

(11) iso ongelma jos se tapahtuu talvella.
big problem if it happens in.winter
'a big problem if it happens in the winter.' (F-657, adult B1)

(12) Alaastella ei saa otta mukaan kouluun,
in.primary.school not get take with to.school
koska sellaiset säännöt.
because such rules
'In primary school, it is not allowed to bring to school because such rules.' (F-200, adolescent A1)

To evaluate the frequency and significance of these problems, the number of sentences without a finite verb was counted. These sentences were found on all proficiency levels, and also in the L1 texts (**Table 4**), although they were most common on the lower proficiency levels in the adult learner data. Other sentences considered problematic were counted after coding the T-units into the data.

Four possible solutions to these clause-level annotation problems were considered. The first of these was to include only sentences containing grammatical clauses. While this decision would solve the problems of clause-level coding of sentences with no finite verb, it would not solve the issues related to the number of clauses within those sentences in which there was a finite verb. It would also mean excluding one fifth of the sentences in the adult learners' texts on the two lowest proficiency levels. Secondly, consideration was given to the possibility of counting the number of clauses based on the number of finite verbs present in the texts (e.g., Verspoor et al., 2017). Although this would provide a solution to the problem of counting the number of clauses within the sentences containing at least one finite verb form, it would be affected by sentences not

Table 4: Number and percentage of sentences containing at least one finite verb, no verb, or at least one non-finite or ambiguous verb form.

Writer group	Finite verb		No verb		Other		Total sentences
	n	%	n	%	n	%	n
Adult A1	256	80	53	16	13	4	322
Adult A2	300	81	67	18	5	1	372
Adult B1	440	94	26	6	3	1	469
Adult B2	366	92	23	6	7	2	396
Adult C1	525	95	21	4	4	1	550
Adult C2	354	93	25	7	3	1	382
Adolescent A1	78	90	4	5	5	6	87
Adolescent A2	136	99	0	0	1	1	137
Adolescent B1	191	98	3	2	1	1	195
L1 year 7	264	92	19	7	4	1	287
L1 year 8	262	95	10	4	4	1	276
L1 year 9	105	95	6	5	0	0	11

containing any finite verbs. The third possible solution was to introduce a new production unit, similar to the sub-clausal element suggested by Foster et al. (2000) for analysing spoken language. While this solution would address issues related to labelling segments without a finite verb, it would introduce two new issues. On the one hand, it would mean that the exact boundaries of these units would become important if one wanted to measure their length or the clause length in words, because all words in these new units would need to be excluded from the word count of the clauses. On the other hand, it would create a need to introduce new measures in which these new units were included. Otherwise, it could entail excluding these new units and their content from the analysis. The fourth solution to the clause-level annotation problems was to also consider segments such as the grammatically incomplete clauses in examples (11) and (12) as attempted clauses and, therefore, to code them as clauses. While this solution would make it possible to include all the data in the analysis with the quantitative measures, it would create segments labelled clauses that do not fall within the original definition, in which a finite verb was required. We will return to this issue in Section 5.

4.4. T-units

A T-unit was defined as a production unit within a sentence consisting of one main clause and all the subordinate clauses connected to it directly or via another subordinate clause. In applying this definition to the data, problems similar to those in segmenting the data into clauses were encountered. First, the use of punctuation created segments in which there seemed to be a sentence boundary within a T-unit, as in example (3). Second, some dependent clauses had a grammatically incomplete clause as their main clause, as in example (11), and some segments beginning with a subordinator were not complete clauses, as in example (12).

Another type of sentence without an apparent main clause was also encountered. In the data, there were sentences that consisted of two clauses, one starting with a subordinator (e.g., *koska* 'because') and the other with a coordinating conjunction (e.g., *tai* 'or'), as in example (13). There were also sentences in which a clause starting with a subordinator seemed to be the main clause of the other clause or clauses in the sentences, as in example (14), in which the clause *Jos ajattelen* 'If I think' seems to be the main clause of two indirect questions rather than a subordinate clause of either of them. With this kind of sentence, analysis of the context is needed to determine the relationship between the clauses.

(13) Koska he eivät saisi olla kauan, tai he
because they not should be for.long or they
eivät saisi surffata nettissä.
not should surf in.net
'Because they should not be for long, or they should not surf the web.' (F-062, adolescent A2)

(14) Jos ajattelen missä Suomi geograafisesti
if I_think where Finland geographically
sijaitsee ja mitä luonnolla on meille
is.located and what nature has us
tarjottavana?
to.offer
'If I think where Finland is geographically located and what nature has to offer us?' (F-420, adult B2)

Sentences containing problems with either the number of clauses or their status as an independent or dependent clause were counted. These sentences were encountered throughout the data on all proficiency levels as well as in the L1 texts. Problematic sentences were more frequent in the adolescent learners' texts (between 22% on level A2

and 9% on B1) than in the adult learners' texts (between 13% on level A1 and 5% on C2), and the problems were not limited to the lower proficiency levels or to isolated texts. Rather, examples were spread across the data, and there was at least one problematic sentence in 40% or more of the L2 texts. There were fewer problematic sentences in the L1 data, but at least one such sentence could be found in 32% of the year 8 students' texts.

To resolve these issues, the use of the sentence as a superordinate unit was reconsidered, as some of the problems could have been solved by coding T-units across perceived sentence boundaries. This would, however, have led to treating some punctuation as erroneous, or ignoring it, which would be problematic, given that in writing, the boundaries of production units cannot be indicated by pauses or intonation, as they can in spoken language. Two other issues to be addressed were the coding of grammatically incomplete clauses or sub-clausal units, and their status as independent or dependent. These problems could have been solved by using an alternative production unit instead of the T-unit, namely the AS-unit, introduced by Foster et al. (2000) for analysing spoken language. While this solution would have acknowledged the sub-clausal units and their role in the superordinate units, it would also have disregarded the sentence boundaries the writer had marked with punctuation.

5. Discussion

When measuring learner language complexity with quantitative measures based on production units such as words, clauses, sentences, and T-units, it is important to split the data into these units reliably and consistently (e.g., Ellis & Barkhuizen, 2005; Pallotti, 2015). Nevertheless, as the results of this study show, learner language texts cannot always be divided into the aforementioned production units without making exceptions or leaving loose ends. In other words, as Rimmer (2006, p. 508) points out, authentic language does not always fit "into neat pigeon holes". It is therefore important to explicitly define the production units used and to make visible the exceptions allowed or the amount of data omitted. This information should always be included when reporting research findings.

In the present study, a sentence was defined as a segment indicated by the writer with punctuation or other orthographic means. As it was marked by the writer, a sentence was considered relevant also to the writer (cf. Peters, 1983). Therefore, it was selected as the superordinate unit (cf. Bardovi-Harlig, 1992; Ellis & Barkhuizen, 2005), and all the texts were first segmented into sentences, which were then split into clauses. In the clause-level annotation, clause boundaries and information on coordination and subordination, including information about the main clause of each dependent clause, were annotated where possible. Unclear cases were analysed and the number of sentences in which they occurred was counted. All of the words were annotated as belonging to a sentence and all sentences were annotated to contain a minimum of one independent clause (and thus also at least one T-unit), even when the sentence did not contain a finite verb or when it began with a subordinator. While these decisions led to segments not falling within the definition of the intended production units, they ensured that all the data were included in every annotation level and that they would be included in quantitative measures of syntactic complexity in future studies using this corpus.

These solutions leave room for criticism. They do, however, resonate with earlier findings of the difficulty of fitting learner language into these production unit categories (e.g., Foster et al., 2000; Rimmer, 2006), and they seem to suggest that reliance on production units that are not necessarily found in learner language could be one of the reasons behind inconsistencies in the results that have been obtained using these measures (e.g., Housen et al., 2019; Ortega, 2003; Wolfe-Quintero et al., 1998). In light of the results and the findings of other studies, three different solutions could be considered. One is forcing learner language into the categories used in quantitative measures, as was done in this study. Another is introducing new units of measure for quantitative research, as, for example, Foster et al. (2000) have done. A third solution is to analyse learner language from a more qualitative perspective and, for example, look for qualitative changes and development in selected linguistic features, as has been done by Reiman (2011) in a study on the development of transitive constructions in written learner Finnish.

There are a number of limitations to this study. The data were split into the production units by one person only. It was therefore impossible to negotiate problematic segments and calculate inter-coder agreement. Comparing the sentence-level results with two other segmentations revealed, however, only minor differences between segmentations in identifying words and sentences, which suggests that the sentence-level coding could be considered reliable enough. On the clause level, the problematic segments and their frequency of occurrence were based on the interpretations of one annotator; another annotator could have made different decisions and arrived at different results. While high inter-annotator agreement enhances the reliability of coding, having more annotators would not have eliminated the need to interpret parts of learner language, to adjust the definitions of production units used, or both.

The target language in this study was Finnish, a morphologically rich language, and it is possible that some of the ambiguities are language-specific. The data used in this study come from a heterogeneous group of learners with different proficiency levels. Some of the segmenting difficulties, such as those related to unsystematic use of punctuation, may also be related to the nature of the data. These issues, nonetheless, should be taken into account when making comparisons between studies within one language or studies on different target languages.

6. Conclusion

The level of detail in learner language coding and in reporting the process naturally depends on the aims and the research questions of each individual study. Nevertheless, segments that are problematic for coding in the data and their potential effect on result, should always

be acknowledged. This is essential for accumulating evidence on the development of complexity and for comparability across studies.

Segments which are problematic for coding could also be seen as potential sources of new information, and they could prove to be worth studying in more detail if a more qualitative approach to investigating complexity was adopted. Analysing the actual structures used by learners instead of forcing all learner language into predefined production unit categories could give new insights into the development of learner language and its complexity.

Notes

[1] The standard Finnish spelling is to separate the number and the unit.

[2] CEFLING = Linguistic basis of the Common European framework for L2 English and L2 Finnish (http://www.jyu.fi/hytk/fi/laitokset/kivi/tutkimus/hankkeet/paattyneet-tutkimushankkeet/cefling).

[3] For challenges in using the same rating scales for L1 and L2 texts, see, for example, Toropainen et al. (2012).

[4] It is available under an open licence at http://turkunlp.github.io/Finnish-dep-parser/. For this study, the branch 'master' updated May 9, 2016 was used.

[5] In Finnish, negation is expressed not with an invariable negation word but with a negation verb (e.g., Karlsson 2015: 82) that agrees with the subject in person and is followed by the main verb (e.g., Lue-n. read-PRS-1SG 'I am reading.', E-n lue. NEG-1SG read 'I am not reading.').

Acknowledgements

I would like to thank Maisa Martin, Jarmo Jantunen, the two anonymous reviewers and the editorial team for their valuable comments.

References

Alanen, R., Huhta, A., & **Tarnanen, M.** (2010). Designing and assessing L2 writing tasks across CEFR proficiency levels. In I. Bartning, M. Martin, & I. Vedder (Eds.), *Communicative proficiency and linguistic development: Intersections between SLA and language testing research* (pp. 21–56). EUROSLA Monographs series, 1. Retrieved from http://eurosla.org/monographs/EM01/EM01home.html

Bardovi-Harlig, K. (1992). A second look at T-unit analysis: Reconsidering the sentence. *TESOL Quarterly, 26*(2), 390–395. DOI: https://doi.org/10.2307/3587016

Biber, D., Gray, B., & **Poonpon, K.** (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly, 45*(1), 5–35. DOI: https://doi.org/10.5054/tq.2011.244483

Biber, D., Johansson, S., Leech, G., Conrad, S., & **Finegan, E.** (1999). *Longman grammar of spoken and written English*. London: Longman.

Booij, G. (2012). *The grammar of words.* Oxford: Oxford University Press.

Brants, T. (2000). Inter-Annotator agreement for a German newspaper corpus. *LREC*. Retrieved from http://www.lrec-conf.org

Brunni, S., Lehto, L., Jantunen, J., & **Airaksinen, V.** (2015). How to annotate morphologically rich learner language. Principles, problems and solutions. *Bergen Language and Linguistics Studies, 6*, 133–152. DOI: https://doi.org/10.15845/bells.v6i0.812

Bulté, B., & **Housen, A.** (2012). Defining and operationalising L2 complexity. In A. Housen, V. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency. Complexity, accuracy and fluency in SLA* (pp. 21–46). Amsterdam and Philadelphia: John Benjamins. DOI: https://doi.org/10.1075/lllt.32.02bul

Byrnes, H, Maxim, H., & **Norris, J.** (2010). Realizing advanced foreign language writing development in collegiate education: Curricular design, pedagogy, assessment. *The Modern Language Journal, 94*(Supplement), 1–235. Retrieved from http://www.jstor.org/stable/40985261. DOI: https://doi.org/10.1111/j.1540-4781.2010.01139.x

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment.* Retrieved from https://rm.coe.int/1680459f97

Crossley, S., & **McNamara, D.** (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing, 26,* 66–79. DOI: https://doi.org/10.1016/j.jslw.2014.09.006

Ellis, R., & **Barkhuizen, G.** (2005). *Analysing learner language.* Oxford: Oxford University Press.

Foster, P., Tonkyn A., & **Wigglesworth G.** (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics, 21*(3), 354–375. DOI: https://doi.org/10.1093/applin/21.3.354

Granger, S. (2002). A bird's-eye view of learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 3–33). Amsterdam: Benjamins. DOI: https://doi.org/10.1075/lllt.6.04gra

Hakulinen, A., & Karlsson, F. (1980). Finnish syntax in text: Methodology and some results of a quantitative study. *Nordic Journal of Linguistics, 3*(2), 93–129. DOI: https://doi.org/10.1017/S0332586500000536

Hakulinen, A., Karlsson, F., & **Vilkuna, M.** (1996). *Suomen tekstilauseiden piirteitä: kvantitatiivinen tutkimus* (2nd ed.). Helsinki: University of Helsinki.

Hakulinen, A., Vilkuna, M., Korhonen, R., Koivisto, V., Heinonen, T. R., & **Alho, I.** (2004). *Iso suomen kielioppi.* Helsinki: Suomalaisen Kirjallisuuden Seura.

Haspelmath, M. (2011). The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica, 45*(1), 31–80. DOI: https://doi.org/10.1515/flin.2011.002

Haverinen, K., Nyblom, J., Viljanen, T., Laippala, V., Kohonen, S., Missilä, A., Ojala, S., Salakoski, T., & **Ginter, F.** (2014). Building the essential resources for Finnish: The Turku Dependency Treebank. *Language Resources and Evaluation, 48*(3), 493–531. DOI: https://doi.org/10.1007/s10579-013-9244-1

Housen, A., De Clercq, B., Kuiken, F., & **Vedder I.** (2019). Multiple approaches to complexity in second language research, *Second Language Research, 35*(1), 3–21. DOI: https://doi.org/10.1177/0267658318809765

Huhta, A., Alanen, R., Tarnanen, M., Martin, M., & **Hirvelä, T.** (2014). Assessing learners' writing skills in a SLA study: Validating the rating process across tasks, scales and languages. *Language Testing, 31*(3), 307–328. DOI: https://doi.org/10.1177/0265532214526176

Kalliokoski, J. (2006). Virke, dialogisuus ja argumentaatio: irralliset sivulauseet ja toisella kielellä kirjoittaminen. In T. Nordlund, T. Onikki-Rantajääskö, T. Suutari, & H. Forsberg (Eds.), *Kohtauspaikkana kieli: näkökulmia persoonaan, muutoksiin ja valintoihin* (pp. 212–231). Helsinki: Suomalaisen Kirjallisuuden Seura.

Karlsson, F. (2015). *Finnish: an essential grammar.* London: Routledge.

Leech, G., & **Svartvik, J.** (2002). *A communicative grammar of English* (3rd ed). London: Longman.

Lieko, A. (1992). *The develoment of complex sentences. A case study of Finnish.* Helsinki: Suomalaisen Kirjallisuuden Seura.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics, 15*(4), 474–496. DOI: https://doi.org/10.1075/ijcl.15.4.02lu

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly, 45*(1), 36–62. DOI: https://doi.org/10.5054/tq.2011.240859

Martin, M. (2013). Sentences and clauses as complexity measures in second language writing: a segmentation experiment. In M. Järventausta, & M. Pantermöller (Eds.), *Finnische Sprache, Literatur und Kultur im deutschsprachigen Raum – Suomen kieli, kirjallisuus ja kulttuuri saksankielisellä alueella* (pp. 185–198). Greifswald: Veröffentlichungen der Societas Uralo-Altaica.

Martin, M., Mustonen, S., Reiman, N., & **Seilonen, M.** (2010). On becoming an independent user. In I. Bartning, M. Martin, & I. Vedder (Eds.), *Communicative proficiency and linguistic development: Intersections between SLA and language testing research* (pp. 57–80). EUROSLA Monographs series, 1. Retrieved from http://eurosla.org/monographs/EM01/EM01home.html

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics, 24*(4), 492–518. DOI: https://doi.org/10.1093/applin/24.4.492

Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research, 31*(1), 117–134. DOI: https://doi.org/10.1177/0267658314536435

Peters, A. (1983). *The units of language acquisition.* Cambridge: Cambridge University Press.

Quirk, R., Greenbaum, S., Leech, G., & **Svartvik, J.** (1972). *A grammar of contemporary English.* London: Longman.

Ragheb, M., & **Dickinson, M.** (2011). Avoiding the comparative fallacy in the annotation of learner corpora. In G. Granena, J. Koeth, S. Lee-Ellis, A. Lukyanchenko, G. P. Botana, & E. Rhoades (Eds.), *Selected proceedings of the 2010 Second Language Research Forum: Reconsidering SLA research, dimensions, and directions* (pp. 114–124). Somerville, MA: Cascadilla Proceedings Project. Retrieved from http://www.lingref.com/cpp/slrf/2010/paper2620.pdf

Rehbein, I., Hirschmann, H., Lüdeling, A., & **Reznicek, M.** (2012). Better tags give better trees – or do they? *Linguistic Issues in Language Technology, 7*(10), 1–18. Retrieved from https://journals.linguisticsociety.org

Reiman, N. (2011). Two faces of complexity: Structural measures and diversity of constructions. *Nordand, 6*(2), 9–23.

Rimmer, W. (2006). Measuring grammatical complexity: The Gordian knot. *Language Testing, 23*(4), 497–519. DOI: https://doi.org/10.1191/0265532206lt339oa

Toropainen, O, Härmälä, M., & **Lahtinen, S.** (2012). Kaksi asteikkoa, kaksi eri tilannetta: äidinkielellä ja vieraalla kielellä kirjoitettujen tekstien kriteeripohjaisen arvioinnin haasteita. *AFinLA-e: Soveltavan kielitieteen tutkimuksia 4*, 60–79. Retrieved from http://journal.fi/afinla/article/view/7038

Verspoor, M., Lowie, W., Chan, H., & **Vahtrick, L.** (2017). Linguistic complexity in second language development: variability and variation at advanced stages. *Recherches en didactique des langues et des cultures, 14*(1), 1–27. DOI: https://doi.org/10.4000/rdlc.1450

Vilkuna, M. (2003). *Suomen lauseopin perusteet.* Helsinki: Edita.

Visapää, L. (2008). *Infinitiivi ja sen infinittiisyys: tutkimus suomen kielen itsenäisistä A-infinitiivikonstruktioista.* Helsinki: Suomalaisen Kirjallisuuden Seura.

Wolfe-Quintero, K., Inagaki, S., & **Kim, H.** (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity. Technical report No. 1.* Honolulu: Second Language Teaching and Curriculum Center.

How to cite this article: Mylläri, T. (2020). Words, clauses, sentences, and T-units in learner language: Precise and objective units of measure? *Journal of the European Second Language Association*, 4(1), 13–23. DOI: https://doi.org/10.22599/jesla.63

Submitted: 22 February 2020 **Accepted:** 20 July 2020 **Published:** 07 August 2020