Relating Production Units and Alignment Units in Translation Activity Data

Michael Carl and Arnt Lykke Jakobsen

Dept. of International Languages Studies & Computational Linguistics
Copenhagen Business School, 2000-Frederiksberg, Denmark

Abstract. The definition and characterisation of Translation Units (TUs) in human translation is controversial and has been described in many different ways. This paper looks at TUs from a translation process perspective: we investigate the sequences of keystrokes which have been typed during translation production and re-define TUs in terms of text production units (PUs). We correlate those units with translation equivalences in the translation product, so-called alignment units (AUs) and compare the translation performance of student and professional translators on a small translation task of 160 words from English into Danish. In contrast to what has frequently been assumed, our data reveals that TUs are rather coarse, as compared to the notion of 'translation atoms', comprising several AUs, and they are particularly coarse for professional translators.

1 Introduction

There is a large body of literature on segmentation in translation which can be separated into two fundamentally different kinds: research into human translation processes seeks to find basic segments of activities in the *translation process*, whereas others think of the segments more statically as properties observable in the *translation product* i.e. correspondences in pairs of texts as a result of a translation process. Accordingly, there is a confusion in the usage of the term "translation unit" (TU), which sometimes refers to the former and sometimes to the latter type of unit. However, it is by no means clear that there is isomorphism between the units that a translator has in mind during translation and the correspondencies which can be made out later in the final translation product.

According to [1] translation units are lexicological units: they are signs, each with de Saussures two components, the *signifiant* and the *signifié*. Such a unit is "the smallest segment of the utterance where the cohesion of signs is such that they cannot be translated separately" [1, p:16]. We will refer to this kind of unit, which can be detected as translation equivalences in the final translation product, as Alignment Units (AUs).

In line with [2, p:14] we adopt the more dynamic view, seeing a TU as "the section of text which the translator focusses on at any one time". Similarly, for [3, p:254] the TU is the "translator's focus of attention at a given time in the translation process". This definition implies that TUs cannot be directly observed in the user activity data (UAD), such as in the translator's keystrokes or gaze movements, but relate to activations in

Carl M. and Lykke Jakobsen A.
Relating Production Units and Alignment Units in Translation Activity Data.
DOI: 10.5220/0003024400370046
In *Proceedings of the 7th International Workshop on Natural Language Processing and Cognitive Science (ICEIS 2010)*, page
ISBN: 978-989-8425-13-3

the translator's mind. However, we assume that there is no appreciable lag between a translator's focus of attention and what he is typing.[1]

It therefore becomes possible to investigate units of translation production and assume that they are related to TUs. We define a production unit (PU) as a sequence of:

1. successive keystrokes in time that are not interrupted by a pause longer that a given PU segmentation threshold. Only deletion and insertion keystrokes are considered. Navigation activities, using the mouse or combinations of keystrokes are ignored.[2]
2. successive keystrokes in text that produce a coherent piece of text. Text producing and deleting activities are part of the same PU only if they are in close proximity to each other.

The boundary of a PU is thus defined to lie between two successive keystrokes that are separated by more than a certain delay in time, or if the second keystroke contributes to production or revision of a different piece of text [5]. Thus, deletions and corrections are possible in one PU if they are within the vicinity of the current cursor position.

Note that not all conceivable PUs are equally likely to reflect a unit of the translator's focus of attention. As is the case for AUs, TUs must also comprise a coherent set of signs that "cannot be translated separately", if, as [6] suggests, skilled translators proceed "little at a time" combining source text (ST) and target text (TT) segments, so that "each segment forms a fragment of bi-text in their minds".

The paper investigates various PU segmentation thresholds to maximise the likelihood of the PU being part of a TU. We expect PUs to satisfy two properties: 1. PUs should represent complete meaning entities rather than arbitrary sequences of keystrokes. 2. PUs should align with AU boundaries, rather than crossing their lines. We are looking for a PU segmentation threshold that maximises these properties.

In section 2 we describe the experimental setup for data acquisition, the hard and software used to collect the UAD as well as details about the translation task. Section 3 looks into details and analyses a data segment in depth. Section 4 applies the analysis technique to a set of 24 translations and provides a large number of PU correlations.

2 Data Acquisition

We base our research on a translation experiment in which 12 professional and 12 student translators produced translations using the version of the Translog [7] software[3] developed as part of the EU Eye-to-IT project[4]. Translog presents the ST in the upper part of the monitor, and the TT is typed in a window in the lower part of the monitor.

[1] Note that this is an adapted version of the "eye-mind assumption" [4], which hypothesises that "there is no appreciable lag between what is being fixated and what is being processed" [4, p:331].

[2] Some people read electronic texts by moving the cursor on the fixated words. If navigation keystrokes were to count in PUs, we would obtain long PUs consisting only of navigation activity.

[3] The keylog portion of the software can be obtained from www.translog.dk

[4] cogs.nbu.bg/eye-to-it/

When the start button is pressed, the ST is displayed and eye movement and keystroke data begins to be registered. The task of the translator is to type the translation in the lower window. After having completed the translation, the subject presses a stop button, and the translation, together with the translation process data, are stored in a log file.

This so-called User Activity Data (UAD) is then transformed into a relational data structure which allows us to map eye movements and keyboard actions onto ST and TT positions and vice versa: [8, 9] describe how keyboard actions are related to the TT words to which they contribute and how TT words are mapped on ST words, so that for (almost) each keystroke, we can determine to which ST translation it contributes. In order to do so automatically, the relational data structure requires, besides the preparation of the keyboard and eye movement data, also the alignment information between the ST and the TT.

For the data set of the 24 translations we semi-automatically aligned the ST and TT and converted the alignment and UAD into the required structure.

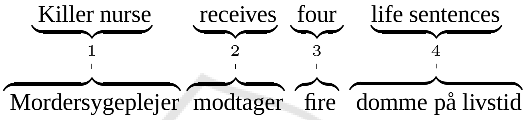

Fig. 1. A segment of the product data shows four AUs aligning 6 ST words and 6 TT words; AUs were manually aligned.

3 Analysis of Production Units

In this section we look more closely at a time segment of 20 seconds in which the translation shown in figure 1 was typed. We develop and discuss properties of the PUs and show how these properties generalise to the entire translation session. In section 4 we apply these criteria to all 24 translations.

Details of the keystroke segmentations are reproduced in tables 1 and 2. During these 20 seconds, 56 keystrokes were produced. Each different PU segmentation threshold groups the data differently with different properties of the produced PUs.

The 400ms segmentation in table 1 groups the 56 keystrokes into 10 PUs. Due to the pauses of 894ms and 488ms after “y” and “j” respectively, the word “Mordersygeplejerske” is segmented into 3 PUs. Note that the suffix “ske” is later deleted and thus does not appear in the final translation in figure 1. On the other hand, the two AUs “modtager” and “fire” are written smoothly so that these two AUs are grouped into one PU. The “start” column in table 1 gives the timestamp when the PU was started to be typed. The “dur” column shows the time needed to complete the PU, and “pause” shows the interval of time following the PU until the next keystroke was processed. Note that the duration is 0 if the PU contains only one keystroke. The “AU” column shows which AUs are generated by each PU. Thus, the first three PUs all contribute to the translation of the first AU_1 while the fourth PU contains AU_2 and AU_3.

A number of corrections occur in the following segments: first the letter “s” is written and then deleted in the next segment. The deletion is indicated in brackets “(s)”.

Table 1. Properties of PUs as generated with 400ms segmentation.

NR	start	dur	pause	type	AU	PU
1	10485	1397	894	SAWU	1	Mordersy
2	12776	944	488	WUWU	1	geplej
3	14208	717	432	WUSA	1	erske
4	15357	2177	2719	SASA	2,3	modtager fire
5	20253	0	702	SAWA	4	s
6	20955	0	4141	WUWU	4	(s)
7	25096	118	886	WUWU	4	li
8	26100	192	669	WUWU	4	(li)
9	26961	1051	536	SASU	4	domme på
10	28548	907	749	SUSA	4	livstid

Then “li” is produced and then deleted “(li)”. Finally the translation “domme på livstid” is typed but segmented into 2 PU due to the delay of 536ms after “på”.

The degree to which a PU coincides with a word in the TT translation or an AU boundary in the ST is indicated by its type. A PU can start and/or end at a word boundary. For instance “livstid” is a complete word and PU$_{10}$ starts and ends at a word separator. “Mordersy” in contrast is the beginning of a word while “erske” is the ending of that word. Accordingly, PU$_1$ starts at a word boundary and PU$_3$ ends with it.

In addition, a word (or segment) in the target language can start and/or end at an AU boundary. For instance, “livstid” is the last part of a compound which is grouped in AU$_4$, and so it ends but does not start at the boundary of AU$_4$. “domme på” is the beginning of that same compound and so PU$_9$ starts, but does not end, at the AU$_4$ boundary.

Thus the type of a PU consists of four positions (bits), each of which can take two values: The first two positions indicate properties for the beginning of the PU, and the last two positions indicate properties of its end. The values indicate whether or not the PU aligns with word boundaries in the target language, and whether or not it aligns with boundaries of the AUs of which it is a translation. These values are represented by the letters [SWAU] which have the following meanings:

S	first/last character of PU was a word separator (space, comma, semicolon, colon) or immediately following a separator.
W	first/last character of PU was not a word separator (and not immediately following a separator)
A	first/last character of PU was at an AU boundary.
U	first/last character of PU was not at an AU boundary.

Thus, “SAWU”, as in the first line of table 1 indicates that the PU “Mordersy” starts at the beginning of a word (S), and it aligns the beginning of an AU (A). The last two letters indicate that this PU ends in the middle of a word (W) and in the middle of an AU (U). Ideally, as discussed in the introduction, a PU should start and end with a word separator and/or an AU boundary, such as “modtager fire” in line 4 of table 1. Segmentations in the middle of a word would (perhaps) indicate that attention is focussed on spelling or typing problems, rather than on translation. Thus, a PU of type “WUWU” (e.g. line 2: “geplej”) indicates that the segment neither starts nor ends at a word or an AU boundary. Such segments provide little insight into the cognitive processes of translators, since they do not coincide with meaning units, as e.g. words and AUs do. However, in the introduction we have argued that PUs should represent signs, with a *signifié*, which is difficult to see in the case of “geplej” or “li”.

Four out of the 10 segments in table 1 are of this type, indicating that a 400ms segmentation threshold does not correspond to the “cognitive” rhythm of segmentation that we are looking for.

Table 2. Segmentation 800ms (above) and 1500ms (below) of the same 56 keystrokes from table 1. The notation “s(s)” and “[s]” are equivalent, meaning typing and deletion of the bracketed expression.

start	dur	pause	type	AU	PU
10485	1397	894	SAWU	1	Mordersy
12776	4758	2719	WUSA	1,2,3	geplejerske modtager fire
20253	702	4141	SAWU	4	s(s)
25096	118	886	WUWU	4	li
26100	4104	3064	WUSA	4	(li)domme på livstid
10485	7049	2719	SASA	1,2,3	Mordersygeplejerske modtager fire
20253	702	4141	SAWU	4	[s]
25096	5108	3064	WUSA	4	[li]domme på livstid

The 800ms PU pattern in table 2 generates only half the number of PUs of those for the 400ms segmentation. Only one (20%) of them is a “WUWU” segment while the remaining four (80%) either start or end with a word separator. This pattern is even more obvious in the 1500ms segmentation of table 2, where all segments show a linguistically plausible beginning or end. On the other hand, segmentation with longer thresholds includes more characters and subsumes more than one AU. Thus, the average lengths of 400ms, 800ms and 1500ms segments of the first 56 keystrokes are 5.6, 11.2 and 18.7 keystrokes, respectively.

Table 3 provides those figures for the entire translation by P13 with various segmentation thresholds for PU patterns. It shows the number of PU segments, their average length in characters (#char), the percentage of linguistically plausible SASA and S.S. segments and the percentage of the implausible WUWU segments. Under the 400ms, 800ms and 1500ms segmentation, translator P13 produced 98, 39 and 23 PUs. The optimum PU segmentation threshold seems to be around 800ms to 1000ms, where a maximum number of segments are linguistically plausible, and at the same time the segments do not comprise too many AUs.

4 Segmentation in Writing

This section investigates various PU segmentation thresholds for all 24 translations. As mentioned earlier in section 2, an English text of 160 words was translated by 12 professional and 12 student translators into Danish. As shown in table 4, there is a large variance in both overall time needed for the translation and in the number of PUs produced. Professionals took on average slightly more than 5 minutes to translate the

Table 3. Number and properties of PUs for translator P13, generated under various segmentation thresholds.

thresh.	#PU	#char	%WUWU	%SASA	%S.S.
200ms	355	2.6	27.04	33.80	37.75
400ms	98	9.4	16.33	37.76	44.90
800ms	39	23.5	2.56	43.59	64.10
1000ms	30	30.5	0.00	50.00	73.33
1500ms	23	39.8	0.00	47.83	69.57
2500ms	14	65.4	0.00	42.86	71.43

text (320 seconds), whereas students took more than 6 minutes (379 seconds). That is, on average professional translators needed 84% of the time needed by students to produce the translation. For both groups there was approximately a factor of 3 between the fastest and slowest translator.

Table 4. Translation time (T-time) and number of PUs for different kinds of segmentation for the 12 professional and 12 student translators.

Professionals		#PU			Students		#PU		
Tr.	T-time	400ms	800ms	1500ms	Tr.	T-time	400ms	800ms	1500ms
P15	170636	70	18	3	S6	228584	107	39	15
P14	209142	105	38	13	S18	260454	133	50	17
P2	258782	96	38	15	S16	285898	125	57	31
P13	265762	98	39	23	S24	322110	158	81	38
P20	281148	162	50	22	S11	350663	131	70	43
P3	316730	138	62	32	S4	353340	174	63	31
P1	349750	107	33	18	S12	374499	127	76	50
P21	352497	154	69	33	S10	377134	111	71	49
P7	362404	141	78	50	S17	411156	132	68	35
P8	379272	114	58	33	S23	425860	176	99	48
P9	389931	145	71	47	S22	507021	170	105	56
P19	510366	177	84	46	S5	654681	218	116	70
av.	320535	126	53	28	av.	379283	147	75	40

Similarly, on average students produced the translations with 30% more PUs than professionals, and there was a factor of almost 4 between the smallest and largest number of PUs produced for both groups. These relations are also plotted in figure 2. The black rectangular symbols in figure 2 indicate the relation between the translation time and the number of segments produced with the 400ms PU segmentation threshold, the triangular symbols those of the 1500ms PU threshold, and yellow squares and diamonds represent student and professional translators, respectively. All translation thresholds indicate a strong correlation between translation time and the number of segments. That is, the more the translation is fragmented into a larger number of segments, the longer is the translation time.[5]

[5] This correlation is not necessary: many fast-typed segments with short pauses in between them could amount to the same overall translation time as few segments with long inter-segmental pauses.

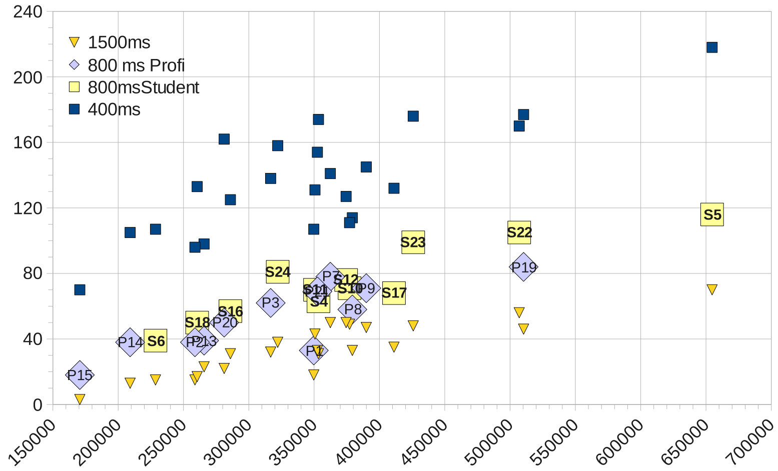

Fig. 2. The graph shows a strong correlation between the translation time (horizontal) and the number of PUs (vertical) for three types of segmentation.

The PUs produced by professionals are, on average, longer and the time needed per PU is (on average) higher for professionals than for those of the students.

Table 5. Average duration and length of PUs with different segmentation thresholds, and the average typing speed (keystrokes per second) for professional and student translators. (Typing averages are high because only within-PU keystroke intervals were counted and because PUs with only one keystroke are recorded with no time duration).

	Professionals			Students		
	400ms	800ms	1500ms	400ms	800ms	1500ms
average PU duration in ms	1001	3113	6896	854	2216	5024
average PU length in chars	7.46	17.61	33.55	6.37	12.54	23.24
average (PU chars/PU dur.)	8.22	5.44	4.83	7.18	5.70	5.02
median (PU chars/PU dur.)	7.15	5.61	4.77	6.97	5.53	4.80

Average duration and length (in characters) for various PU segmentation thresholds is given in table 5. Depending on the PU thresholds, students generate, on average, 15% to 30% shorter segments in length as well as in duration. They also produce more segments than professional translators, as we have already shown above in table 4 and in figure 2.

However, the typing speed in terms of characters per time within each PU does not seem to vary much between professionals and students. The values in table 5 indicate that average and median inner-segment typing speed between successive keystrokes decreases with a growing PU threshold. For instance, at 800ms segmentation threshold, professionals produce 5.44 characters, while students produce 5.7 keystrokes per second. The typing behaviour of students is more fragmented than that of professionals, with more pauses longer than the segmentation threshold, but when the segments are actually typed, the speed with which successive keystrokes are produced seems to be identical for both groups.[6]

With a segmentation threshold of 1500ms the longest PU was produced by translator P15, with 183 characters. With a threshold of 800ms the longest PU was produced by translator P14, comprising 18 AUs and the following 153 characters[7]:

> fordi Norris ikke kunne lide at arbejde med [g]ældre mennesker. Alle hans ofte
> var s[v][a]vagelige ældre kvin[ger]der med hjerteproblemer. ALle ville [8]

This TU starts with a subordinate clause, that is, the second half of a sentence, it then contains a whole sentence and ends with the beginning of a third sentence at word position 150.

Table 6. Number and properties of PUs for different segmentation thresholds: SASA: PUs start and end with a word separator and an AU boundary, S.S.: PUs start and end with a word separator, WUWU: PUs start and end in the middle of a word.

	200ms	400ms	800ms	1000ms	1500ms	2500ms
#PU	8060	3293	1557	1245	842	524
%SASA	11.39	37.16	48.96	49.72	41.92	37.40
%S.S.	24.54	45.68	53.64	54.22	55.11	50.0
%WUWU	36.72	20.77	7.32	6.27	4.75	4.96

A more detailed listing of the types of PU is given in table 6. The table shows that the type and distribution of PUs change under different PU thresholds. It gives an overview of the number of PUs produced with 6 different thresholds and shows the percentage of linguistically more and less meaningful segments, similar to table 3. The table does not make a distinction betwee students and professional translators. With increasing segmentation time, the number of generated segments decreases, and the percentage of meaningful segments increases. A dramatic change of this effect can be observed up to ca. 800ms: the meaningless “WUWU” segments fall below 10% and the linguistically coherent ones grow beyond 50%. Beyond this margin, values change less quickly.

5 Conclusions

The paper investigates activity data of student and professional translators: An English text of 160 words was translated by 12 professional and 12 student translators into Danish.[9] All keystrokes and eye movements were recorded using the Translog software.

[6] Note that the typing speed was computed as the lapse of time between two (or more) successive keystrokes, so that a PU consisting of a single keystroke would count as 0ms duration. This explains the relatively fast typing speed.

[7] deletions are in square brackets. The notation “[ger]” means that first “ger ” was produced (including blank) and then deleted.

[8] The source text of this passage is: “that Norris disliked working with old people. All of his victims were old weak women with heart problems. All of them could”

[9] The ST is reproduced in the Appendix.

We investigate and correlate three types of unit: production units (PUs) of keystrokes where no two keystrokes are separated by a pause longer than a given threshold, alignment units (AUs) which reflect translation equivalences in the translation product, and translation units (TUs) which represent the translators' focus of attention.

The goal was 1) to determine a threshold for the segmentation of PUs such that they most likely conform to criteria of TUs, and 2) to investigate properties of PUs for students and professional translators.

Various thresholds of maximal delay between successive keystrokes are explored to group sequences of keystrokes into PUs. PUs are considered intelligible if their boundaries coincide maximally with linguistic (i.e. word) boundaries in the target language and with the boundaries of translation atoms as defined by AUs. Our investigation shows that pauses in writing activity of approximately 1000ms length produce segments of maximal linguistic plausibility, and are, thus, indicative of cognitive processing units. That is, a new TU is likely to have started if a pause of 1 second or more can be observed with no keystroke.

Our findings can be summarised as follows:

- The number of PUs correlates strongly with translation time: the longer the translation time, the more fragmented is the translation into segments (figure 2).
- Professional translators produce translations more quickly than students (table 4).
- Professional translators produce longer PUs than students in terms of time, as well as in terms of the number of characters (table 5).
- Professional and student translators type PUs at approximately the same speed (table 5).
- Longer PUs coincide better with word boundaries than shorter PUs (table 6)

Only a small percentage of PUs coincide with a single AU. Rather than producing a minimal unit or a "translation atom", we find that translators produce maximal segments, which seem to increase with the capacity and training of the translator. In a further study, we intend to investigate properties of these segments in more depth so as to construct an inventory of cognitive operations associated with the PUs.

References

1. Vinay, J.P., Darbelnet, J.: Stylistique comparée du français et de l'anglais. Didier, Paris (1958)
2. Bennett, P.: Translation Units in Human and Machine. Babel 40 (1994) 12–20
3. Alves, F., Vale, D.C.: Probing the Unit of Translationin time: aspects of the design and development of a web application for storing, annotating and querying translation process data. Across Language and Cultures 10 (2009) 251–273
4. Just, Carpenter, P.: A theory of reading: from eye fixations to comprehension. Psychological Review 87 (1980) 329–354
5. Carl, M., Kay, M., Kristian T.H., J.: Long Distance Revisions in Drafting and Post-editing. In: Cicling, Iasi, Romania (2010)
6. Harris, B.: Bi-text: A New Concept in Translation Theory. Language Monthly 54 (1988) 8–10
7. Jakobsen, A.L.: Logging target text production with Translog. In: [10]. (1999) 9–20

8. Carl, M.: Triangulating product and process data: quantifying alignment units with keystroke data. Copenhagen Studies in Language 38 (2009) 225–247
9. Carl, M., Jakobsen, A.L.: Towards statistical modelling of translators activity data. International Journal of Speech Technology 12 (2010) http://www.springerlink.com/content/3745875x22883306/.
10. Hansen, G., ed.: Probing the process in translation: methods and results. Volume 24 of Copenhagen Studies in Language. Copenhagen: Samfundslitteratur (1999)

Appendix: Source Test

```
Killer nurse receives four life sentences
Hospital Nurse Colin Norris was imprisoned for life today for
the killing of four of his patients. 32 year old Norris from
Glasgow killed the four women in 2002 by giving them large
amounts of sleeping medicine. Yesterday, he was found guilty
of four counts of murder following a long trial. He was given
four life sentences, one for each of the killings. He will have
to serve at least 30 years. Police officer Chris Gregg said
that Norris had been acting strangely around the hospital. Only
the awareness of other hospital staff put a stop to him and to
the killings. The police have learned that the motive for the
killings was that Norris disliked working with old people. All
of his victims were old weak women with heart problems. All of
them could be considered a burden to hospital staff.
```