Integrating Treebank Annotation and User Activity in Translation Research

Michael Carl and Henrik Høeg Müller

Abstract The Center for Innovation of Translation and Translation Technology (CRITT) environment at Copenhagen Business School (CBS) draws on primarily two types of NLP resources, namely treebanks and the logging of user activity data (UAD) during text production and translation activities, in order to do research into the cognitive processes that lie behind translation activity. In this paper we make a short presentation of the Copenhagen Dependency Treebank (CDT), and elaborate how UAD is obtained and represented in Translog-II. Finally, the paper discusses some general perspectives on how process-oriented translation research methodology could benefit from the integration of UAD with structural linguistic information in the form of linguistically annotated text data.

1 Introduction

The main focus of the CRITT (Center for Innovation of Translation and Translation Technology) research environment is on the empirical and experimental study of translation processes with an applied, technological aim. Our research designs involve data elicitation methods (keystroke logging) and behavioural measuring technologies (eyetracking), as well as parallel linguistically annotated text collections (treebanks). To record user activity data (UAD), CRITT has developed the computer programme Translog-II, which logs keystrokes, mouse activities, and gaze movements during text production. With respect to treebanks, the CRITT translation research programme has devised the Copenhagen Dependency Treebank (CDT), an NLP resource which provides information about language structure and meaning on various levels. While the CDT annotates the static structure of the parallel, trans-

Michael Carl · Henrik Høeg Müller
Copenhagen Business School, Languages & Computational Linguistics,
Frederiksberg, Denmark
e-mail: \{mc.isv, hhm.isv\}@cbs.dk

lated texts, Translog-II provides information on how the parallel data was actually created during the translation process.

Since around the mid 90s most texts are generated by humans using a (computer) keyboard, but still there is hardly any empirical data available that is suited to investigate how translations are generated, and to uncover and describe the processes by which humans produce translations. A central aim of CRITT is to overcome this gap. In this paper, we seek to connect the two worlds of product and process annotation. The paper is structured as follows. In Section 2, it is illustrated, very briefly, how linguistic structure is annotated in CDT, including alignment. In Section 3, focus is on how UAD is structured in Translog-II. Section 4 offers some speculations about the possible benefits derived from integrating CDT and Translog-II, while in section 5 we comment on the predictability of the translators' behaviour. Finally, section 6 sums up the central points.

2 The Copenhagen Dependency Treebank

The Copenhagen Dependency Treebank (CDT) (Trautner-Kromann, 2003) is a multilingual open NLP resource which consists of linguistically annotated parallel text collections of approx. 60.000 words each for Danish, English, German, Italian, and Spanish. The CDT is based on a unified dependency annotation which includes not only syntax, but also fine-grained analyses of morphological, discourse, and anaphoric structure. Moreover, in order to extent its applicability potential to MT, the resource has an alignment system of translational equivalences that allows us to specify relations between words or word groups in the source and target language that correspond to each other with respect to meaning or function (Buch-Kromann et al, 2009).

Figure 1 shows the primary dependency tree for the sentence "Kate is working to earn money" (top arrows), enhanced with secondary subject relations (bottom arrows). The arrows point from governor to dependent, with the relation name written at the arrow tip.

CDT also allow for morphological annotation which deals with derivation and composition. The internal structure of words is encoded as a dependency tree through an operator notation scheme (Müller and Durst-Andersen, 2012).

Figure 2 illustrates how anaphora (bottom arrows) and discourse (top arrows) are annotated in CDT (Korzen and Müller, 2011). The arrows indicating anaphoric relations run from the antecedent to the anaphora, while the discourse arrows go from the top node of the governing segment to the top node of the dependent segment, the top node being typically, but not necessarily, the verb. Figure 3 plots the alignment of a Danish-English translation together with their syntactic annotation.

This multilevel annotation distinguishes CDT from other treebank projects which tend to focus on a single linguistic level, and it has the advantage of not obliging us to limit the kind of linguistic relations that can be annotated, and not having to draw precise, and often arbitrary, boundaries between morphology, syntax, and discourse.

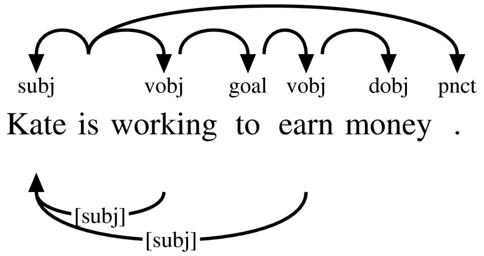

Fig. 1 Primary dependency tree (top) augmented with secondary subject dependency relations (bottom).

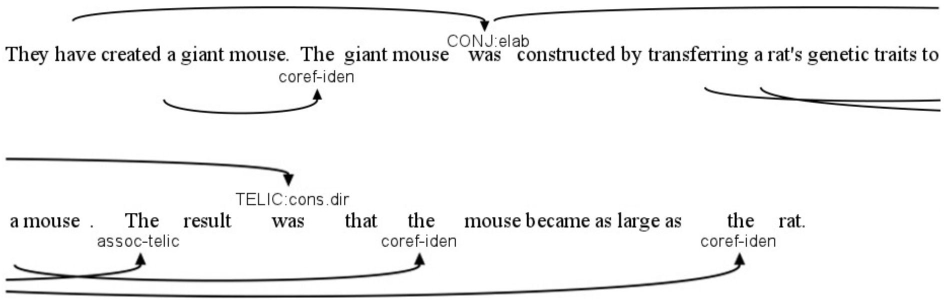
CONJ:elab
They have created a giant mouse. The giant mouse was constructed by transferring a rat's genetic traits to
coref-iden
TELIC:cons.dir
a mouse . The result was that the mouse became as large as the rat.
assoc-telic
coref-iden
coref-iden

Fig. 2 Discourse and anaphoric relations in the CDT on the bottom and top lines respectively. The annotation schema is independent from the morphological, syntactic, and alignment information.

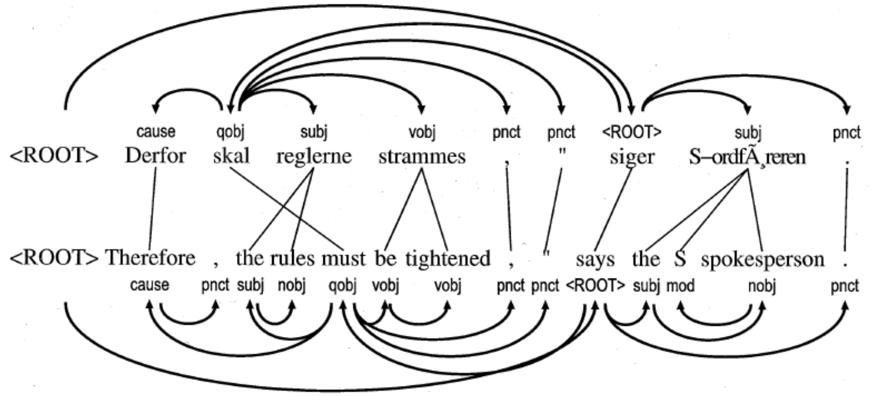
cause
qobj
subj
vobj
pnct
pnct
<ROOT>
subj
pnct
<ROOT> Derfor skal reglerne strammes , " siger S–ordfÃ¸reren .
<ROOT> Therefore , the rules must be tightened , " says the S spokesperson .
cause
pnct
subj
nobj
qobj
vobj
vobj
pnct
pnct
<ROOT>
subj
mod
nobj
pnct

Fig. 3 The Figure shows an example of an alignment between Danish and English. For the sake of clarity only the syntactic annotation appears in the figure.

3 The Structure of Translation Process Data

The UAD acquisition software Translog-II (Jakobsen, 1999) logs keystrokes and mouse activities during text production and it also records gaze movements on the texts when connected to an eye tracker. While Translog-II can also be used for reading and writing research, we present here the logging protocol when used as a translation tool. Translog-II divides the screen horizontally into a source text window (top) and a target text window (bottom) in which the translator types her translation. The activity data is collected during a translation session for later analysis and can be replayed in a replay mode (Carl et al, 2011; Carl and Dragsted, 2011; Dragsted, 2010).

The process data also consist of three resources, the gaze sample points, fixations and keystroke information. Gaze sample points consist of screen coordinates looked at by the left and right eye, as well as pupil dilation at a particular time. In addition, the gaze data contains the windows (source or target) in which the gaze (i.e. the left eye sample) was detected. The log file also contains the location of the closest character (i.e. the cursor position of the character in the ST or TT) to the gaze sample point, and the location of that character on the screen.

Fixations are computed based on sequences of gaze samples. Fixations group together a number of near-distance eye-gaze samples which represent a time segment in which a word (or symbol) is fixated. In our current representation, fixations have a starting time, a duration, and a cursor position which refers to an index in the ST or TT.

Translog-II also logs a number of keyboard and mouse activities. We distinguish five different types of keystrokes: insertion, deletion, editing, navigation and the return key. Each of the permitted keystrokes has a slightly different representation, depending on whether or not the keystroke is visible on the screen and whether it involves a deletion or not. All insertion keys have a cursor position and screen position (X/Y, Width and Height). There are two different kinds of explicit deletions, the `[Back]` and the `[Delete]` key. Navigation keystrokes (up, down, left, right, home, end) can be combined with the `Ctrl` key. Navigation keys can also be combined with the `Shift` key, which results in selecting and marking of the sequence.

Translog-II stores the full information in the log file which makes it possible to re-produce all text modifying operations of a translation session outside the data acquisition software without loss of information. In addition to the text and gaze activities, Translog-II also logs interface and system events, such a changing the size of the window, interruption of the translation session, etc. The Translog data has been used in a number of studies (Schou et al, 2009), and we expect that many more investigation strands can be developed, particularly when combining the process data with the annotation of product data.

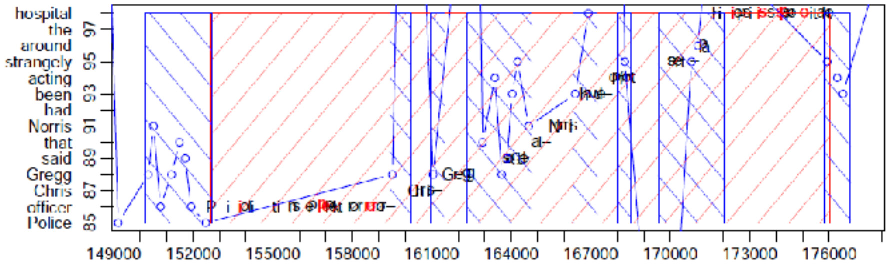

Fig. 4 Translation progression graph showing parallel reading and text production

4 Perspectives of Integrating Translog-II with CDT

In this last part of the paper, we suggest some theoretical considerations as to what could be gained from integrating Translog-II UAD with dependency annotated texts in CDT with respect to conducting basic empirical research into the cognitive processes during translation activities. In general, we expect that the combination of Tranlog-II and CDT data will facilitate enquiries into isomorphisms between, on the one hand, pauses and gaze location/duration and, on the other, linguistic layers in its broadest sense, which could reveal new insights about source text decoding, memory retrieval and encoding of textual segments, and their linguistic structure. The eye-mind hypothesis predicts a strong correlation between gaze location and mind activity, i.e. between what we are looking at and what we are thinking about. In Translation Process Research it is customary to part from the assumption that gaze location reflects the focus of attention of the translator so that longer gaze durations on the ST or TT correspond to bigger decoding (comprehension) and encoding (productions) problems, respectively, presupposing a stratificational process model of translation (Hyrskykari, 2006; Pavlovic and Jensen, 2009).

In Figure 4 we present a graph which illustrates how Translog-II records gaze data and keystrokes in a translation process (Carl and Dragsted, 2011). It presents a translation progression graph for the translation of a English source sentence into Danish:

- English source sentence:
  Police officer Chris Gregg said that Norris had been acting strangely around the hospital
- Danish translation:
  P[i]olitiins[ep]pekt[rør]ør Chris Gregg sagte at Norris havde opført sig sært på h[i]o[p[si]s]sp[o]italet

The vertical axis represents the ST and the horizontal axis the time span of 28 seconds (149-177) during which the translation took place. The red upwards hatched boxes are clusters of coherent writing activity, whereas the blue circles indicate ST fixations and the corresponding blue downwards hatched boxes symbolize fixation units. Figure 4 shows that fixations are not equally distributed over the ST. There are two large fixations units in the progression graph, i.e. between seconds 149 and 152, and between 162 and 165. In the first case, the translator's eyes moved around in the text chunk "Police officer Chris Gregg said that Norris" before beginning the new sentence, whereas in the second case the reading activity of "Gregg said that Norris had been acting strangely" occurs while typing the translation.

In the future, CRITT intends to devise computational mechanisms to integrate the process data (as shown in figure 4) with CDT information, i.e. morphological, syntactic, discourse and anaphoric, as well as alignment information as shown in figures 1 to 3.

5 Predictability of Translators' Behaviour

The distribution of gaze points and keystrokes of a single translator, of course, reflects an individual pattern of translation challenges. In order to investigate whether different translators translating the same text face similar difficulties at the same text positions, Carl and Dragsted (2011) conducted an experiment where they looked into the amount of ST gaze activity that was detected before a translation was typed, and compared the gaze durations of 5 translators. The result is plotted in Figure 5.

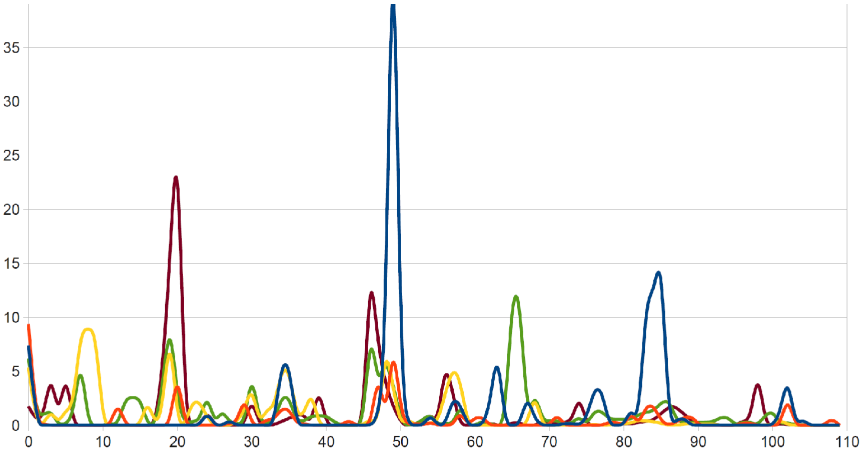

Fig. 5 Relative amount of fixation duration before/during production of the translation of ST words

The horizontal axis enumerates words in the ST. The vertical axis plots the amount of ST gaze time before typing the translation of the ST word, where each color represents a translator.

Figure 5 shows that the translators spent gaze time on close ST fragments, indicating that they have similar problems at this point in the translation.

6 Conclusion

The central problem is, however, that UAD must be directly correlated with an analysis of linguistic data in a systematic way to formulate hypotheses about the problems faced by the translators, and to establish links between sensory-motor clues and linguistic characteristics of the translated expressions. With the words of Angelone (2010) “non-articulated indicators, such as pauses and eye-fixations, give us no real clue as to how and where to allocate the uncertainty”. However, if UAD is directly connected to a parallel treebank with multilevel linguistic annotation, the possibilities of systematically analyzing correlations between gaze fixations and keystrokes and underlying linguistic structure of the texts are promising to uncover the translation processes. Specifically, we assume that the integration of these two NLP resources would allow us to correlate patterns of UAD with patterns of morphological, syntactic or discourse structure. By mapping dynamic UAD on to structural treebank annotation data, behavioral factors become grounded in linguistics, and in this way we may gain a better understanding of the interconnection between text production and comprehension processes, i.e. of how cognitive activity does (or does not) correspond with the linguistic categories and complexity we are used to deal with from an analytical and theoretical perspective. The integration of product and process data opens the possibility to investigate the extent to which patterns of UAD are related to standard linguistic units in the form of phrases, sentences, etc. or whether there are correspondences between eye-gaze regression patterns and anaphoric paths, discourse structures or complex phrases.

References

Angelone E (2010) Uncertainty, uncertainty management and metacognitive problem solving in the translation task. Translation and Cognition pp 17–40

Buch-Kromann M, I K, Müller HH (2009) Uncovering the lost structure of translations with parallel treebanks. Copenhagen Studies in Language 38:199–224

Carl M, Dragsted B (2011) Inside the monitor model: Processes of default and challenged translation production. In: Contrastive Linguistics, Translation Studies, Machine Translation – What can we Learn from Each Other? workshop held in conjunction with the Conference of the German Society for Computational Linguistics and Language Technology (GSCL 2011), Hamburg, Germany

Carl M, Dragsted B, Lykke Jakobsen A (2011) On the systematicity of human translation processes. In: Proceedings of Translation Careers and Technologies: Convergence Points for the Future (Tralogy 2011), Paris, France

Dragsted B (2010) Coordination of reading and writing processes in translation. Translation and Cognition, American Translators Association Scholarly Monograph Series, Benjamins, Amsterdam/Philadelphia

Hyrskykari A (2006) Utilizing eye movements: Overcoming inaccuracy while tracking the focus of attention during reading. Special issue: Attention aware systems

Jakobsen A (1999) Logging target text production with translog. In: Hansen G (ed) Probing the process in translation: methods and results, Copenhagen Studies in Language, vol 24, Samfundslitteratur, Copenhagen, pp 9–20

Korzen I, Müller HH (2011) The copenhagen dependency treebank. Forskellige niveauer samme relationer

Müller HH, Durst-Andersen P (eds) (2012) Ny forskning i Grammatik. Odense Universitetsforlag, Odense

Pavlovic N, Jensen KTH (2009) Eye tracking translation directionality. In: Pym A, Perekrestenko A (eds) Translation Research Projects 2, Intercultural Studies Group, Tarragona, pp 93–109, URL \url{http://isg.urv.es/publicity/isg/publications/trp_2_2009/index.htm}

Schou L, B D, Carl M (2009) Ten years of translog. Copenhagen Studies in Language 37:37–51

Trautner-Kromann M (2003) The Danish dependency treebank and the DTAG treebank tool. In: 2nd Workshop on Treebanks and Linguistic Theories, Växjo, Sweden, pp 217–220