From Character to Word Level: Enabling the Linguistic Analyses of Inputlog Process Data

Mariëlle Leijten
Flanders Research Foundation
University of Antwerp
Department of Management
Belgium
marielle.leijten@ua.ac.be

Lieve Macken
LT^3, Language and Translation Technology Team, University College Ghent and Ghent University
Belgium
lieve.macken@hogent.be

Veronique Hoste
LT^3, Language and Translation Technology Team, University College Ghent and Ghent University
Belgium
veronique.hoste@hogent.be

Eric Van Horenbeeck
University of Antwerp
Department of Management
Belgium
eric.vanhorenbeeck@ua.ac.be

Luuk Van Waes
University of Antwerp
Department of Management
Belgium
luuk.vanwaes@ua.ac.be

Abstract

Keystroke-logging tools are widely used in writing process research. These applications are designed to capture each character and mouse movement as isolated events as an indicator of cognitive processes. The current research project explores the possibilities of aggregating the logged process data from the letter level (keystroke) to the word level by merging them with existing lexica and using NLP tools. Linking writing process data to lexica and using NLP tools enables researchers to analyze the data on a higher, more complex level.

In this project the output data of Inputlog are segmented on the sentence level and then tokenized. However, by definition writing process data do not always represent clean and grammatical text. Coping with this problem was one of the main challenges in the current project. Therefore, a parser has been developed that extracts three types of data from the S-notation: word-level revisions, deleted fragments, and the final writing product. The within-word typing errors are identified and excluded from further analyses. At this stage the Inputlog process data are enriched with the following linguistic information: part-of-speech tags, lemmas, chunks, syllable boundaries and word frequencies.

1 Introduction

Keystroke-logging is a popular method in writing research (Sullivan & Lindgren, 2006) to study the underlying cognitive processes (Berninger, 2012). Various keystroke-logging programs have been developed, each with a different focus[1]. The programs differ in the events that are logged

[1] A detailed overview of available keystroke logging programs can be found on http://www.writingpro.eu/logging_programs.php.

Proceedings of the EACL 2012 Workshop on Computational Linguistics and Writing, pages 1–8,
Avignon, France, April 23, 2012.

(keyboard and/or mouse, speech recognition), in the environment that is logged (a program-specific text editor, MS Word or all Windows-based applications), in their combination with other logging tools (e.g., eye tracking and usability tools like Morae) and the analytic detail of the output files. Examples of keystroke-logging tools are:

- Scriptlog: Text editor, Eyetracking (Strömqvist, Holmqvist, Johansson, Karlsson, & Wengelin, 2006),
- Inputlog: Windows environment, speech recognition (Leijten & Van Waes, 2006),
- Translog: Text editor, integration of dictionaries (Jakobsen, 2006) (Wengelin et al., 2009).

Keystroke loggers' data output is mainly based on capturing each character and mouse movement as isolated events. In the current research project[2] we explore the possibilities of aggregating the logged process data from the letter level (keystroke) to the word level by merging them with existing lexica and using NLP tools.

Linking writing process data to lexica and using NLP tools enables us to analyze the data on a higher, more complex level. By doing so we would like to stimulate interdisciplinary research, and relate findings in the domain of writing research to other domains (e.g., Pragmatics, CALL, Translation studies, Psycholinguistics).

We argue that the enriched process data combined with temporal information (time stamps, action times and pauses) will further facilitate the analysis of the logged data and address innovative research questions. For instance, *Is there a developmental shift in the pausing behaviors of writers related to word classes, e.g., before adjectives as opposed to before nouns (cf. cognitive development in language production)? Do translation segments correspond to linguistic units (e.g., comparing speech recognition and keyboarding)? Which linguistic shifts characterize substitutions as a sub type of revisions (e.g., linguistic categories, frequency)?*

A more elaborate example of a research question in which the linguistic information has added value is: *Is the text prodcution of causal markers more cognitive demanding than the production of temporal markers*? In reading research, evidence is found that it takes readers longer to process sentences or paragraphs that contain causal markers than temporal markers. Does the same hold for the production of these linguistic markers? Based on the linguistic information added to the writing process data researchers are now able to easily select causal and temporal markers and compare the process data from various perspectives *(cf. step 4 - linguistic analyses)*.

The work described in this paper is based on the output of Inputlog[3], but it can also be applied to the output of other keystroke-logging programs. To promote more linguistically-oriented writing process research, Inputlog aggregates the logged process data from the character level (keystroke) to the word level. In a subsequent step, we use various Natural Language Processing (NLP) tools to further annotate the logged process data with different kinds of linguistic information: part-of-speech tags, lemmata, chunk boundaries, syllable boundaries, and word frequency.

The remainder of this paper is structured as follows. Section 2 describes the output of Inputlog, and section 3 describes an intermediate level of analysis. Section 4 describes the flow of the linguistic analyses and the various linguistic annotations. Section 5 wraps up with some concluding remarks and suggestions for future research.

2 Inputlog

Inputlog is a word-processor independent keystroke-logging program that not only registers keystrokes, mouse movements, clicks and pauses in MS Word, but also in any other Windows-based software applications.

Keystroke-logging programs store the complete sequence of keyboard and/or mouse events in chronological order. Figure 1 represents "*Volgend jaar*" *('*Next Year*')* at the character and mouse action level.

The keyboard strokes, mouse movements, and mouse clicks are represented in a readable output for each action (e.g., 'SPACE' refers to the spacebar, LEFT Click is a left mouse click, and 'Movement' is a synthesized representation of a continuous mouse movement). Additionally, timestamps indicate when keys are pressed and released, and when mouse movements are made. For each keystroke in MSWord the position of

[2] FWO-Merging writing process data with lexica - 2009-2012

[3] http://www.inputlog.net/

the character in the document is represented as well as the total length of the document at that specific moment. This enables researchers to take the non-linearity of the writing process into account, which is the result of the execution of revisions during the text production.

Event Type	Output	Position	Doclength	StartTime	EndTime	ActionTime	PauseTime
focus	Twitter3.docm - Microsoft Word			3604	3604	0	3604
mouse	LEFT Click			6428	6677	249	6428
mouse	Movement			9594	10577	983	2917
mouse	Movement			23244	24024	780	12667
mouse	LEFT Click			24118	24212	94	94
mouse	Movement			24134	24134	0	0
mouse	Movement			24258	24290	32	124
keyboard	V	0	0	26864	26973	375	2574
keyboard	o	1	1	27160	27238	78	296
keyboard	l	2	2	27363	27534	171	203
keyboard	g	3	3	27456	27581	125	93
keyboard	e	4	4	27534	27706	172	78
keyboard	n	5	5	27675	27784	109	141
keyboard	d	6	6	27862	28018	156	187
keyboard	SPACE	7	7	27987	28127	140	125
keyboard	j	8	8	28127	28268	141	140
keyboard	a	9	9	28268	28330	62	141
keyboard	a	10	10	28408	28517	109	140
keyboard	r	11	11	28486	28658	172	78
keyboard	SPACE	12	12	28611	28736	125	125
keyboard	o	13	13	28689	28829	140	78

Figure 1 Example of general analysis Inputlog.

To represent the non-linearity of the writing process the S-notation is used. The S-notation (Kollberg & Severinson Eklundh, 2002) contains information about the revision types (insertion or deletion), the order of the revisions and the place in the text where the writing process was interrupted. The S-notation can be automatically generated from the keystroke-logging data and has become a standard in the representation of the non-linearity in writing processes.

Figure 2 shows an example of the S-notation. The text is taken from an experiment with master students Multilingual Professional Communication who were asked to write a (Dutch) tweet about a conference (VWEC). The S-notation shows the final product and the process needed.

Volgend·jaar·organiseert·{#|4}^{3}VWEC·een·{boeiend·|9}^{8}congres·[over·']^{1}|1[met·als·thema|10]^{9}{over}^{10}·'Corporate·Communication{'|8}^{7}.[.]^{2}|2[·Wat·levert·het·op?'.|7]^{6}·Blijf·[ons·volgen·op|5]^{4}{op·de·hoogte·via|6}^{5}·www.vwec2012.be.|3·

Figure 2. Example of S-notation.

The following conventions are used in S-notation:

- $|_i$: a break in the writing process with sequential number *i*;
- {insertion}$_i$: an insertion occurring after break *i*;
- [deletion]$_i$: a deletion occurring after break *i*.

The example in Figure 2 can be read as follows:

The writer formulates in one segment "*Volgend jaar organiseert VWEC een congres over*" (*'*Next year VWEC organises a conference on*')*. She decides to delete "*over*" (index 1) and then adds the remainder of her first draft "*met als thema 'Corporate Communication. Wat levert het op'?.*" (*'*themed 'Corporate Communication. What is in it for us'?.*')* She deletes a full stop and ends with "*Blijf ons volgen op www.vwec2012.be.*" (*'*Follow us on www.vwec2012.be*')*. The third revision is the addition of the hashtag before VWEC. Then she rephrases "*ons volgen op*" into "*op de hoogte via.*" She notices that her tweet is too long (max. 140 characters) and she decides to delete the subtitle of the conference. She adds the adjective "*boeiend*" ('interesting') to conference and ends by deleting "*met als thema*" ('themed').

3 Intermediate level

At the intermediate level, Inputlog data can also be used to analyze data at the digraph level, for instance, to study interkey intervals (or digraph latency) in relation to typing speed, keyboard efficiency of touch typists and others, dyslexia and keyboard fluency, biometric verification etc. For this type of research, logging data can be leveled up to an intermediate level in which two consecutive events are treated as a unit (e.g., un-ni-it).

Grabowski's research on the internal structure of students' keyboard skills in different writing tasks is a case in point (Grabowski, 2008). He studied whether there are patterns of overall keyboard behavior and whether such patterns are stable across different (copying) tasks. Across tasks, typing speed turned out to be the most stable characteristic of a keyboard user. Another example is the work by Nottbush and his colleagues. Focusing on linguistic aspects of interkey intervals, their research (Nottbusch, 2010; Sahel, Nottbusch, Grimm, & Weingarten, 2008) shows that the syllable boundaries within words have an effect on the temporal keystroke succession. Syllable boundaries lead to increased interkey intervals at the digraph level.

In recent research Inputlog data has also been used to analyze typing errors at this level (Van Waes & Leijten, 2010). As will be demonstrated in the next section, typing errors complicate the analysis of logging data at the word and sentence level because the linear reconstruction is disrupted. For this purpose a large experimental corpus based on a controlled copying task was

analyzed, focusing on five digraphs with different characteristics (frequency, keyboard distribution, left-right coordination). The results of a multilevel analysis show that there is no correlation between the frequency of a digraph and the chance that a typing error occurs. However, typing errors show a limited variation: pressing the adjacent key explains more than 40% of the errors, both for touch typists and others; the chance that a typing error is made is related to the characteristics of the digraph, and the individual typing style. Moreover, the median pausing time preceding a typing error tends to be longer than the median interkey transitions of the intended digraph typed correctly. These results illustrate that further research should make it possible to identify and isolate typing errors in logged process data and build an algorithm to filter them during data preparation. This would benefit parsing at a later stage (see section 4).

4 Flow of linguistic analyses

As explained above, writing process data gathered via the traditional keystroke-logging tools are represented at the *character level* and produce *non-linear data* (containing sentence fragments, unfinished sentences/words and spelling errors). These two characteristics are the main obstacles that we need to cope with to analyze writing process data on a higher level. In this section we explain the flow of the linguistic analyses.

4.1 Step 1 - aggregate letter to word level

Natural Language Processing tools, such as part-of-speech taggers, lemmatizers and chunkers are trained on (completed) sentences and words. Therefore, to use the standard NLP tools to enrich the process data with linguistic information, in a first step, words, word groups, and sentences are extracted from the process data.

The S-notation was used as a basis to further segment the data into sentences and tokenize them. A dedicated sentence segmenting and tokenizer module was developed to conduct this process. This dedicated module can cope with the specific S-notation annotations such as insertion, deletion and break markers.

4.2 Step 2 – parsing the S-notation

As mentioned before, standard NLP tools are designed to work with clean, grammatically correct text. We thus decided to treat word-level revisions differently than higher-level revisions and to distinguish deleted fragments from the final writing product.

We developed a parser that extracts three types of data from the S-notation: word-level revisions, deleted fragments, and the final writing product. The word-level revisions can be extracted from the S-notation by retaining all words with word-internal square or curly brackets (see excerpt 1).

(1 - word level revision)

Delet[r]ion incorrect: Deletrion; correct: deletion
In{s}ertion incorrect: Inertion; correct: insertion

Conceptually, the deleted fragments can be extracted from the S-notation by retaining only the words and phrases that are surrounded by word-external square brackets (2); and the final product data can be obtained by deleting everything in between square brackets from the S-notation. In practice, the situation is more complicated as insertions and deletions can be nested.

An example of the three different data types extracted from the S-notation is presented in the excerpt below. To facilitate the readability of the resulting data, the indices are omitted (3).

(2 - deleted fragments)

Volgend·jaar·organiseert·{#}VWEC·een·{boeiend·}congres·[over·'][met·als·thema]{over}·'Corporate·Communication{'}.[.][·Wat·levert·het·op?'.]·Blijf·[ons·volgen·op]{op·de·hoogte·via|}·www.vwec2012.be.|·

(3 - final writing product)

Volgend·jaar·organiseert·{#}VWEC·een·{boeiend·}congres·[over·'][met·als·thema]{over}·'Corporate·Communication{'}.[.][·Wat·levert·het·op?'.]·Blijf·[ons·volgen·op]{op·de·hoogte·via|}·www.vwec2012.be.|·

English translation
Next year #VWEC organises an interesting conference about Corporate Communication. Follow us on www.vwec2012.be

In sum, the output of Inputlog data is segmented in sentences and tokenized. The S-notation is divided into three types of revisions

and the within-word typing errors are excluded from further analyses.

Although the set-up of the Inputlog extension is largely language-independent, the NLP tools used are language-dependent. As proof-of-concept, we provide evidence from English and Dutch (See Figure 3).

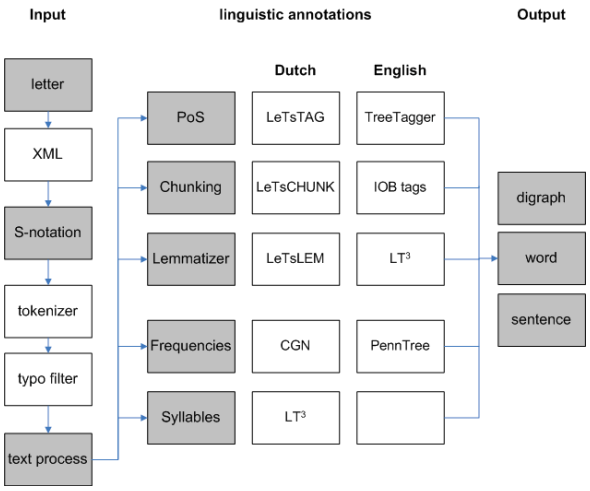

Figure 3 Flow of the linguistic analyses.

4.3 Step 3 – enriching process data with linguistic information

As standard NLP tools are trained on clean data, these tools are not suited for processing input containing spelling errors. Therefore, we only enrich *the final product data* and the *deleted fragments* with different kinds of linguistic annotations. As part-of-speech taggers typically use the surrounding local context to determine the proper part-of-speech tag for a given word (typically a window of two to three words and/or tags is used), the deletions in context are extracted from the S-notation to be processed by the part-of-speech tagger. The deleted fragments in context consist of the whole text string without the insertions and are only used to optimize the results of the linguistic annotation.

(4 - deleted fragments in context)

Volgend·jaar·organiseert·{#}VWEC·een·{boeiend·}congres·[over·'][met·als·thema]{over}·'Corporate·Communication{'}.[.][·Wat·levert·het·op?'.]·Blijf·[ons·volgen·op]{op·de·hoogte·via|}·www.vwec2012.be.|·

For the shallow linguistic analysis, we used the LT^3 shallow parsing tools suite consisting of:

- a part-of-speech tagger (LeTsTAG),
- a lemmatizer (LeTsLEMM), and
- a chunker (LeTsCHUNK).

The LT3 tools are platform-independent and hence run on Windows.

Part of speech tags

The English PoS tagger uses the Penn Treebank tag set, which contains 45 distinct tags. The Dutch part-of-speech tagger uses the CGN tag set codes (Van Eynde, Zavrel, & Daelemans, 2000), which is characterized by a high level of granularity. Apart from the word class, the CGN tag set codes a wide range of morpho-syntactic features as attributes to the word class. In total, 316 distinct tags are discerned.

Lemmata

During lemmatization, for each orthographic token, the base form (lemma) is generated. For verbs, the base form is the infinitive; for most other words, this base form is the stem, i.e., the word form without inflectional affixes. The lemmatizers make use of the predicted PoS codes to disambiguate ambiguous word forms, e.g., Dutch "*landen*" can be an infinitive (base form "*landen*") or plural form of a noun (base form "*land*"). The lemmatizers were trained on the English and Dutch parts of the Celex lexical database respectively (Baayen, Piepenbrock, & van Rijn, 1993).

Chunks

During text chunking syntactically related consecutive words are combined into non-overlapping, non-recursive chunks on the basis of a fairly superficial analysis. The chunks are represented by means of IOB-tags.

In the IOB-tagging scheme, each token belongs to one of the following three types: I (inside), O (outside) and B (begin); the B- en I-tags are followed by the chunk type, e.g., B-VP, I-VP. We adapted the IOB-tagging scheme and added end tag (E) to explicitly mark the end of a chunk. Accuracy sores of part-of-speech taggers and lemmatizers typically fluctuate around 97% to 98%; accuracy scores of 95% to 96% are obtained for chunking.

After annotation, the final writing product, deleted fragments, and word-level corrections are aligned and the indices are restored. Figures 4 and 5 show how we enriched the logged process data with different kinds of linguistic information: lemmata, part-of-speech tags, and chunk boundaries.

We further added some word-level annotations on the final writing product and the deletions,

# revisions	index (begin revision)	index (end revisions)	product	word level corrections	lemma	PoS	Chunk	Syllables	Absolute freq
			Volgend		volgend	ADJ	B-NP	vol-gend	44
			jaar		jaar	N-s	E-NP	jaar	200634
			organiseert		organiseren	V-fin	B-VP	or-ga-ni-seert	13803
1	3	3	#VWEC	_3_[#]_3_VWEC	#VWEC	N-prop-s	B-NP	v-wec	/
			een		een	DET	B-NP	een	2150389
1	8	8	_8_{boeiend}_8_		boeiend	ADJ	I-NP	boe-iend	47
			congres		congres	N-s	E-NP	con-gres	2840
	1								
1	9	1							
1		9							
1	10	10	_10_{over}_10_		over	PREP	B-PP	o-ver	146623
			'Corporate		'Corporate	N-prop	I-PP	cor-po-ra-te	37
1	7	7	Communication'	Communication_7_{'}_7_	communication'	N-prop	E-PP	com-mu-ni-ca-t-io	22
			.		.	PCT			/
1	2	2							
	6								
1		6							
			Blijf		blijven	V-fin	B-VP	blijf	58219
	4								
1		4							
	5		_5_{op		op	PREP	B-PP	op	881849
			de		de	DET	I-PP	de	5827958
			hoogte		hoogte	N-s	E-PP	hoog-te	12674
1		5	via}_5_		via	PREP	I-PP	vi-a	32887
			www.vwec2012.be.		www.vwec2012.be.	SYM	E-PP	www-v-wec-be	/

Figure 4 Final writing product and word-level revisions enriched with linguistic information.

viz., syllable boundaries and word frequencies (see last two columns in Figures 4 and 5).

Syllable boundaries:

The syllabification tools were trained on Celex (`http://lt3.hogent.be/en/tools/timbl-syllabification`). Syllabification was approached as a classification task: a large instance base of syllabified data is presented to a classification algorithm, which automatically learns from it the patterns needed to syllabify unseen data. Accuracy scores for syllabification reside in the range of 92% to 95%.

Word Frequency

Frequency lists for Dutch and English were compiled on the basis of Wikipedia pages, which were extracted from the XML dump of the Dutch and English Wikipedia of December 2011. We used the Wikipedia Extractor developed by Medialab[4] to extract the text from the wiki files. The Wikipedia text files were further tokenized and enriched with part-of-speech tags and

product	deletions	lemma	PoS	Chunk	Syllables	Absolute freq
Volgend						
jaar						
organiseert						
#VWEC						
een						
_8_{boeiend}_8_						
congres						
	_1_[over	over	PREP	B-PP	o-ver	146623
	']_1__9_[met	'met	N-s	B-PP	met	706784
	als	als	PREP	I-PP	als	231748
	thema]_9_	thema	N-s	E-PP	the-ma	5022
_10_{over}_10_						
'Corporate						
Communication'						
.						
	_2_[.]_2_	.	PCT			
	_6_[Wat	wat	PRON-int	B-NP	wat	61958
	levert	leveren	V-fin	B-VP	le-vert	15043
	het	het	DET	B-NP	het	2301736
	op?'.]_6_	op?'.	N-s	E-VP	op	881849
Blijf						
	_4_[ons	ons	PRON	B-NP	ons	8779
	volgen	volgen	V-inf	E-VP	vol-gen	66786
	op]_4_	op	PREP	B-PP	op	881849
_5_{op						
de						
hoogte						
via}_5_						
www.vwec2012.be.						

Figure 5 Deleted fragments enriched with linguistic information.

[4] http://medialab.di.unipi.it/wiki/Wikipedia_Extractor

lemmata. The Wikipedia frequency lists can thus group different word forms belonging to one lemma.

The current version of the Dutch frequency list has been compiled on the basis of nearly 100 million tokens coming from 395,673 Wikipedia pages, which is almost half of the Dutch Wikipedia dump of December 2011.

Frequencies are presented as absolute frequencies.

4.4 Step 4 - combining process data with linguistic information

In a final step we combine the process data with the linguistic information. Based on the time information provided by Inputlog, researchers can calculate various measures, e.g., length of a pause within, before and after lemmata, part-of-speech tags, and at chunk boundaries.

As an example Table 1 shows the mean pausing time before and after the adjectives and nouns in the tweet. Of course, this is a very small-scale example, but it shows the possibilities of exploring writing process data from a linguistic perspective.

	mean pause before	mean pause after	mean pause within
ADJ	1880	671	148
NOUN	728	1455	232
B (begin)	1412	1174	164
E (end)	685	1353	148
I (inside)	730	1034	144

Table 1. Example of process data and linguistic information

In this example the mean pausing time before adjectives is twice as long as before nouns. The pausing time after such a segment shows the opposite proportion. Also pauses in the beginning of chunks are more than twice as long as in the middle of a chunk.

5 Future research

In this paper we presented how writing process data can be enriched with linguistic information. The annotated output facilitates the linguistic analysis of the logged data and provides a valuable basis for more linguistically-oriented writing process research. We hope that this perspective will further enrich writing process research.

5.1 Additional annotations and analyses

In a first phase we only focused on English and Dutch, but the method can be easily applied to other languages as well provided that the linguistic tools are available for a Windows platform.

For the moment, the linguistic annotations are limited to part-of-speech tags, lemmata, chunk information, syllabification, and word frequency information, but can be extended, e.g., by n-gram frequencies to capture collocations.

By aggregating the logged process data from the character level (keystroke) to the word level, general statistics (e.g., total number of deleted or inserted words, pause length before nouns preceded by an adjective or not) can be generated easily from the output of Inputlog as well.

5.2 Technical flow of Inputlog & linguistic tools

At this point Inputlog is a standalone program that needs to be installed on the same local machine that is used to produce the texts. This makes sense as long as the heaviest part of the work is the logging of a writing process. However, extending the scope from a character based analysis device to a system that supplements fine-grained production and process information to various NLP tools is a compelling reason to rethink the overall architecture of the software.

It is not feasible to install the necessary linguistic software with its accompanying databases on every device. By decoupling the capturing part from the analytics a research group will have a better view on the use of its hard- and software resources while also allowing to solve potential copyright issues. Inputlog is now pragmatically Windows-based, but with the new architecture any tool on any OS will be capable to exchange data and results. It will be possible to add an NLP module that receives Inputlog data through a communication layer. A workflow procedure then presents the data in order to the different NLP packages and collects the final output. Because all data traffic is done with XML files, cooperation between software with different creeds becomes conceivable. Finally, the module has an administration utility handling the necessary user authentication and permits.

Acknowledgements

This study is partially funded by a research grant of the Flanders Research Foundation (FWO 2009-2012).

References

Baayen, R. H., R. Piepenbrock, & H. van Rijn. (1993). The CELEX lexical database on CD-ROM. Philadelphia, PA: Linguistic Data Consortium.

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1993). The CELEX lexical database on CD-ROM. Philadelphia, PA: Linguistic Data Consortium.

Berninger, V. (2012). Past, Present, and Future Contributions of Cognitive Writing Research to Cognitive Psychology: Taylor and Francis.

Grabowski, J. (2008). The internal structure of university students' keyboard skills. Journal of Writing Research, 1(1), 27-52.

Jakobsen, A. L. (2006). Translog: Research methods in translation. In K. P. H. Sullivan & E. Lindgren (Eds.), Computer Keystroke Logging and Writing: Methods and Applications (pp. 95-105). Oxford: Elsevier.

Kollberg, P., & Severinson Eklundh, K. (2002). Studying writers' revising patterns with S-notation analysis. In T. Olive & C. M. Levy (Eds.), Contemporary Tools and Techniques for Studying Writing (pp. 89-104). Dordrecht: Kluwer Academic Publishers.

Leijten, M., & Van Waes, L. (2006). Inputlog: New Perspectives on the Logging of On-Line Writing. In K. P. H. Sullivan & E. Lindgren (Eds.), Computer Keystroke Logging and Writing: Methods and Applications (pp. 73-94). Oxford: Elsevier.

Nottbusch, G. (2010). Grammatical planning, execution, and control in written sentence production. Reading and Writing, 23(7), 777-801.

Sahel, S., Nottbusch, G., Grimm, A., & Weingarten, R. (2008). Written production of German compounds: Effects of lexical frequency and semantic transparency. Written Language and Literacy, 11(2), 211-228.

Strömqvist, S., Holmqvist, K., Johansson, V., Karlsson, H., & Wengelin, A. (2006). What keystroke logging can reveal about writing. In K. P. H. Sullivan & E. Lindgren (Eds.), Computer Keystroke Logging and Writing: Methods and Applications (pp. 45-71). Oxford: Elsevier.

Sullivan, K. P. H., & Lindgren, E. (2006). Computer Key-Stroke Logging and Writing. Oxford: Elsevier Science.

Van Eynde, F., Zavrel, J., & Daelemans, W. (2000). Part of Speech Tagging and Lemmatisation for the Spoken Dutch Corpus. Paper presented at the Proceedings of the second International Conference on Language Resources and Evaluation (LREC), Athens, Greece.

Van Waes, L., & Leijten, M. (2010). The dynamics of typing errors in text production. Paper presented at the SIG Writing 2010, 12th International Conference of the Earli Special Interest Group on Writing, Heidelberg.

Wengelin, A., Torrance, M., Holmqvist, K., Simpson, S., Galbraith, D., Johansson, V., & Johansson, R. (2009). Combined eyetracking and keystroke-logging methods for studying cognitive processes in text production. Behavior Research Methods, 41(2), 337-351.