


Investigating Keylogs as Time-Stamped Graphemics


Nicolas Ballier, Erin Pacquetet & Taylor Arnold


Abstract. This article investigates keystroke data, in an attempt to articulate the microlevel of the graphemic level with the macrolevel of text structures. Analyzing the time-stamps of keylogs, we suggest a hierarchy of constituents inspired by speech data and focus on the interaction of graphemic structure, phonological structure and textual structure within the dimension of time. We present the prototype of an R package designed to analyze keylog capture data, taking into account graphemic structures, syllable counts and parsing. Our R package under development offers functions that can be used to analyze the various levels of graphemic constituents produced by typists, from syllable counts to n -gram analysis.

1. Introduction

Current keyboards used with computers have reproduced mechanical and then electric keyboard layout (the QWERTY layout), even though alternative models such as BEPO or EWOPY (Bellis, 2017) have been developed now that the layout of keys on the keyboard is no longer dependent on the physical interactions of keys before hitting the ribbon. Typing is an emerging form of language production that has become part of our everyday lives in modern western societies. Most people use typing to write every day whether it is for professional or personal reasons. There is thus a need to better understand the processes involved in typing through a linguistic perspective, and it is interesting to consider

Nicolas Ballier  0000-0003-2179-1043
Université de Paris, CLILLAC-ARP, F-75013 Paris, France

Erin Pacquetet  0000-0001-9664-8167
Department of Linguistics, 609 Baldy Hall, University at Buffalo, North Campus, Buffalo, NY 14260-6420

Taylor Arnold  0000-0003-0576-0669
Department of Math & Computer Science, 212 Jepson Hall, 221 Richmond Way, University of Richmond, VA 23173

Y. Haralambous (Ed.), *Graphemics in the 21st Century. Brest, June 13-15, 2018. Proceedings Grapholinguistics and Its Applications* (ISSN: 2534-5192), Vol. 1.
Fluxus Editions, Brest, 2019, p. 353–365. <https://doi.org/10.36824/2018-graf-ball>
ISBN: 978-2-9570549-0-9, e-ISBN: 978-2-9570549-1-6

what linguistic information is encoded in typing. Moreover, typed language displays features from both traditional writing and oral speech, as well as features that are very specific to the typing medium and that do not have equivalents in other forms of language production.

Whenever a computer user types on a keyboard, it is possible to collect the timed typing information through keyloggers. The collection of user keystrokes on computers dates back to the beginning of personal computers and is still being used today for many of different purposes.

Keyloggers come both in hardware and software form (although nowadays, there are almost exclusively in software form) and are often devised as spywares that are hidden from the user's awareness. Historically, they have been used both by hackers wanting to recover passwords and by institutions trying to improve password and authentication security by learning individual typing patterns to discriminate between users (Giot, El-Abed, and Rosenberger, 2009), in particular to authenticate users in on-line courses.

Linguistically, keystroke logging is interesting because it enables researchers to witness the timed production of a typed text in a discreet and non-intrusive fashion and in a potentially naturalistic setting. Moreover, this technique requires easily accessible equipment (a computer and a keyboard). This is thus a very practical and accessible way of recording language production. Keystroke logging thus opens the door to not only investigate what language is produced but, and most importantly, how complex linguistic units are constructed and what the underlying processes are (Cislaru and Olive, 2018).

This article presents research in the making on the constituents of keylog capture data. The time stamps of the keys hit when we write texts have mostly been used to perform user authentication (Bergadano, Gunetti, and Picardi, 2002) and some datasets have been produced specifically for this aim, focusing on password typing (Giot, El-Abed, and Rosenberger, 2009; Giot, Ninassi, El-Abed, and Rosenberger, 2012). The past few years have seen an increase in the use of keystroke logging techniques in many areas of academic research. More recently, some studies have begun to address the linguistic data per se, whether to question non-canonical data (Plank, 2016) or to analyze the accelerations in typing (Van Waes, Leijten, and Neuwirth 2006; Leijten and Van Waes 2013). Some constituents have been investigated, either at the word level (Weingarten, Nottbusch, and Will, 2004) or above the word (Chukharev-Khudilaynen 2014; Cislaru and Olive 2016; Cislaru and Olive 2017), for exemple in synchronous computer-mediated communication (Charoenchaikorn, 2019) or in note-taking tasks (Malekian et al., 2019).

In the next section, we analyze the keystroke logs of English-speaking typists writing short essays in examination conditions (Charles C. Tappert, Cha, Villani, and Zack, 2012). We aim to characterize the

flow of typed data in terms of the size of the constituents processed by the typist (*processing chunks*). The aim is to establish the thresholds (and maxima) of relevant pauses to identify the constituents of typed texts, based on the model of the analysis of the prosodic constituents for the prosodic hierarchy (Nespor and Vogel, 2007). We compare, depending on the pauses identified, the span of typed sequences and their number of syllables; we explore possible constraints on the number of syllables cognitively treated for each identified constituent.

2. Datasets

As keystroke logging was primarily developed for spying and hacking, many tools available for keystroke logging are actually spyware. This is of course not desirable for academic research for ethical reasons as the logging has to be confined to the task presented to the test-takers and should stop when the experiment is over. Moreover, most spyware focuses rather on capturing the text typed than the timestamps of typing for the goal is to steal information and not to analyze typing patterns. Therefore, the kind of data collected and the way it is presented when using spyware is not suited to academic research. For instance, most spyware will collect the keys pressed and the timestamp of the typing session, but not the timepresses of each individual keys which are useful when looking at a production from a linguistic point of view.

In order to collect data in a safer and more controlled environment, several keylogging software packages have been designed specifically for research purposes. Among them, we can cite Inputlog (Leijten and Van Waes, 2013) which has been devised specifically for collecting keystrokes in an academic environment. Inputlog is a local keylogger that works with the software Microsoft Word. Once launched, the software opens a word document in which test takers can type freely. The keystrokes and mouse movements are recorded and saved. The software also performs analyses on the data and has a replay tool to re-watch the production. Inputlog is thus very well suited for academic research, but presents some limitations in the required setup for data collection as it has to be performed locally and there is little control over the parameters when using the predefined metrics.

Other tools are devised as sorts of hybrid solutions to collect data online and in an invisible way, but confined to a learning platform. An example of such a keylogger is the Moodle plugin BioAuth devised by Vincent Monaco (Stewart, Monaco, Cha, and Charles C. Tappert, 2011). Moodle is an open source learning management system that many universities worldwide use for course management. The BioAuth plugin enables course administrators to record keystrokes from students who are answering online quizzes hosted on the platform. The collection is lim-

ited to quizzes and stops once the answer is submitted—this makes it safe for students to use. The data is then stored on the Moodle database and can be accessed by the course administrator but not by the student. Students are identified on the platform and each production can be traced back to their typist. The platform also allows to embed various media types within the quiz question, which means that a wide variety of tasks such as picture description or guided production can be performed by the student.

When it comes to academic research, there are a lot of different datasets available for keystroke logging. Since research on the matter is fairly recent and there are no standards for keyloggers and/or experiment protocols, each research question calls for a different dataset and many studies end up collecting their own data, tailored to their needs. Therefore, there is a real plurality in terms of what is available.

In general, we can separate keystroke logging datasets into two categories: long-input and short-input datasets. Long-input datasets are made of long text input, usually answers to a question of at least one sentence. Examples of such studies include work on identifying typists based on stylometry and keystroke features (keystrokes dynamics-based user authentication) (Stewart, Monaco, Cha, and Charles C. Tappert 2011; Monaco, Stewart, Cha, and Charles C Tappert 2013; Kang and Cho 2015). These datasets make it possible to carry out linguistic analysis of the different language units and their mutual interactions.

Short-input datasets typically display typing sequences of a word or less. The most popular types of short-input studies are password studies where researchers attempt to gather information on how a specific typist types a specific password and use machine learning algorithms or biometrics to identify the typist and thus increase protection of accounts and personal data (Giot, El-Abed, and Rosenberger 2009; Killourhy and Maxion 2009). There is however little to no linguistic interest to such datasets as it is made out of very little language and passwords are often constructed as random sequences of characters.

Several datasets have been recently made publicly available. We will show how typing skills can be assessed by copying tasks and we will detail some of the resulting datasets. We use a long text input dataset of college examination answers presented in Charles C. Tappert, Cha, Villani, and Zack (2012). The keystrokes were collected from “40 students of a spreadsheet modelling course in the business school of a four-year liberal arts college” (ibid.). Although the test was administered online, the students did meet in a desktop classroom for each session, providing a controlled environment for the experiment. Tests were taken on Dell keyboards and desktops and the test takers got the opportunity to train on these keyboards beforehand. The test takers were not aware that their keystrokes were being captured at the time of the test. This is therefore a relatively natural setting for keystroke collection. The students took

four online tests of 10 questions each, with a two-week interval between each test. In the dataset, each test taker was assigned a number. A number is also assigned to each session. We used the keystrokes of 38 users, from user 2 to user 43 (users 10, 16, 20 and 36 are not part of the original dataset because they failed to complete the examination). For each key that was pressed by a given user during a given session, the dataset provides timestamps corresponding to the time at which the key was pressed and the time at which it was released. In addition, we are also given the keyname and the JavaScript keycode of each typed key. Table 1 is a sample from the dataset.

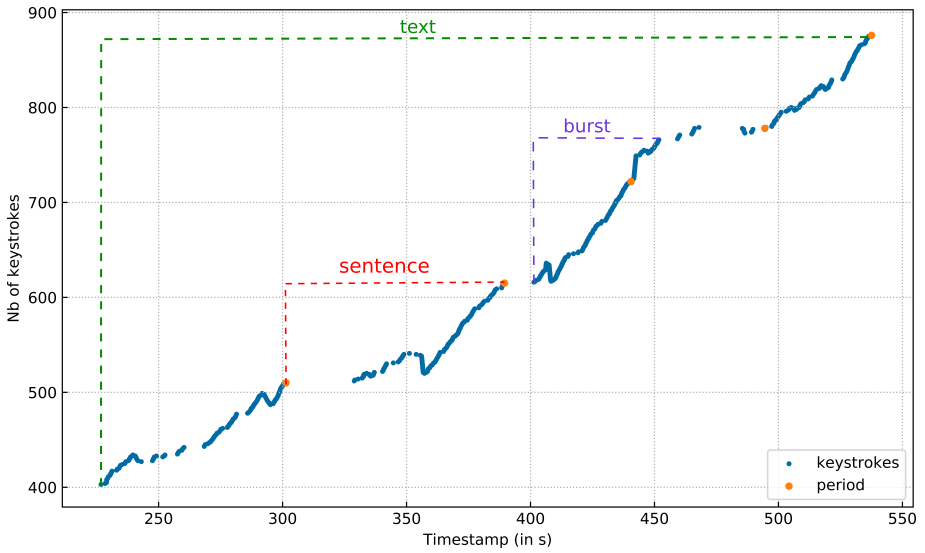
TABLE 1. Sample from the original dataset (Charles C. Tappert, Cha, Villani, and Zack, 2012)

user	session	timepress	timerelease	keycode	keyname
23	14	1301579856926	1301579857102	73	i
23	14	1301579857070	1301579857246	78	n
23	14	1301579857262	1301579857422	32	space
23	14	1301579858302	1301579858462	83	s
23	14	1301579858462	1301579858558	79	o
23	14	1301579858622	1301579858750	76	l
23	14	1301579858990	1301579859086	86	v
23	14	1301579859070	1301579859214	73	i
23	14	1301579859182	1301579859294	78	n
23	14	1301579859262	1301579859374	71	g
23	14	1301579859358	1301579859470	32	space

3. The Prosodic Hierarchy

Using our R package, we can reconstruct texts from this initial input and discuss the clustering of graphemes into higher constituents, whether at syllable, word or chunk level. Our research question can be summed up with Figure 1, which describes how chunks of graphemes (top) can cluster according to the prosodic hierarchy acknowledged in Nespor and Vogel (2007) (bottom) and whose lower constituents were tentatively described for keylogs (Weingarten, Nottbusch, and Will, 2004).

The *constituent model of written word production* (ibid.) distinguishes a graphemic word (W), some lexical constituents (LC) here aptly illustrated by the German compound *Flaschenöffner* (‘bottle opener’), syllables (S) and their phonological sub-constituents (O is the onset, R is for the rhyme), its graphemic layer (G_C stands for the consonant grapheme and G_{Cn} is a ‘consonant grapheme with n letters’ and G_V a vowel grapheme).



Prosodic domains	Syntactic units	Keyboard units
Speech	Text	Text
Paratone (\mathbb{T})	Paragraph	Paragraph
Phonological / prosodic utterance (PU)	(Utterance)	Chunk
Intonational Phrases (I or IP)	Sentence	
Phonological phrase (Φ or PP)	Clause Phrase Heavy NP	
Clitic group (C)	Noun Phrase	
Phonological / prosodic word (ω)		
Foot (Φ or F)	Word	Word
Syllable (σ)		
Mora (μ)		
Segments (phonemes)	Letter	Character

FIGURE 1. Mapping the series of bursts of a writer (top) to the hierarchical structure of the prosodic hierarchy (bottom)

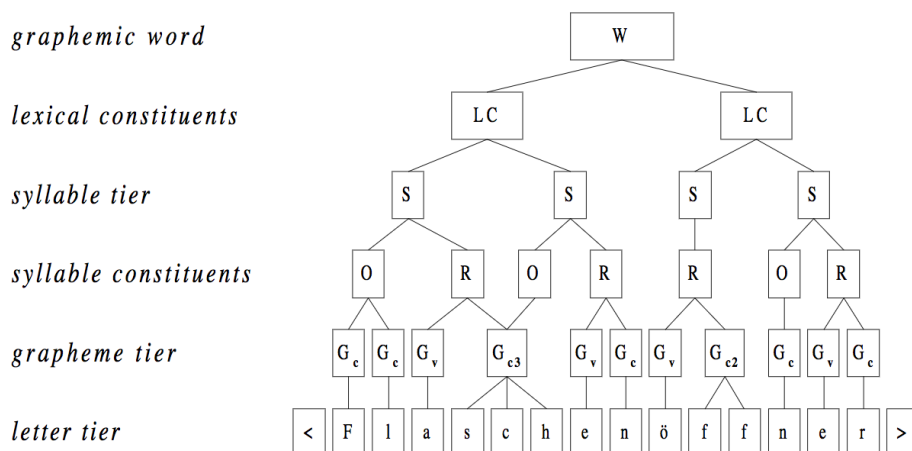


FIGURE 2. The 'constituent model of written word production' (after Weingarten, Nottbusch, and Will 2004)

The last tier corresponds to the letters (see Figure 2 and the presentation of the prosodic hierarchy in Evertz in this volume). It should be noted that in this representation, the final consonants of the syllable, or codas, are not represented and ambisyllabicity is assumed for the consonant represented by the grapheme <sch>, which is both in onset and in rhyme position. Applying this representation to English, this begs the relevance of the graphemic level, taking into account a (sub)constituent such as <th>. Do writers type this grapheme faster in final, medial or initial position (therefore, with different phonological status) and does its morphemic status in *tenth* or *length* have any bearing on the variable performances?

This also questions the status of affixes. The words generated in the dataset were parsed for the presence of strings of words that are commonly defined as 'suffixes' at the end of the words. It is noticeable that words with suffix *-like* endings are typed faster than words without these endings (see the general comparison of speed for words with and without suffixes on Figure 3). These words also have a higher average frequency and a larger number of characters, the latter being usually an indication that the strings are typed more slowly.

4. Backspace Management, Parsing and the Dynamics of Keylog

Above the word, the analysis of writing systems and their representation in written communication needs to take into account the process

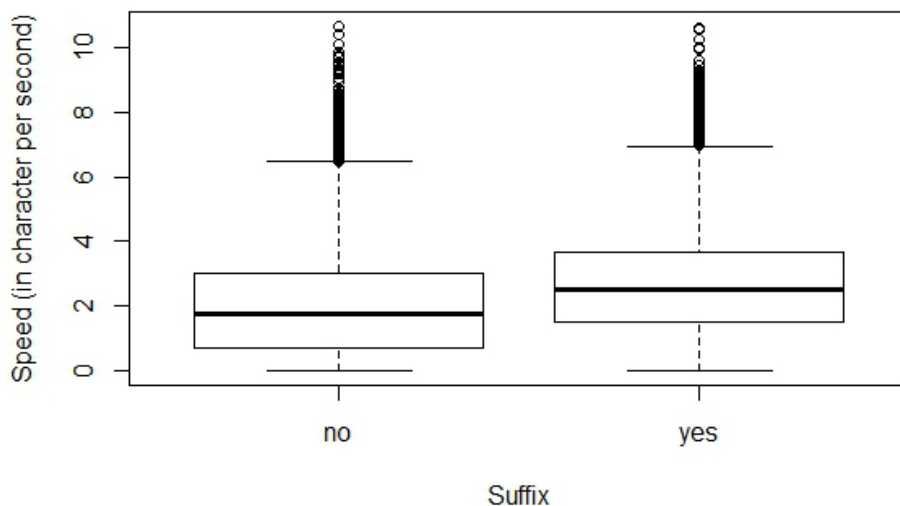


FIGURE 3. Boxplot of the speed of words with and without suffixes

TABLE 2. Frequency and length of words with and without suffixes

	No Suffix	Suffix
Average word frequency	21.09026	10.84965
Average number of characters	3.173415	5.853220

of writing and the complex interaction of revisions and corrections. The last section of the paper will show the benefits of our R scripts to compare the resulting texts and the dynamic processes of typing, especially the use of the backspace key. The dual nature of typed texts is summed up by Mahlow (2015) who advocated the need to address both “the product, i.e., the text where the error is visible for a reader, and the process, i.e., the editing operations causing this error.” We briefly illustrate graphs of inserted letters and repairs (backspace) and the resulting textual structures. As evidenced in the graphs below, we believe the ‘backspace’ key should be granted a special status it may erase complete textual bursts (right) so that we advocate a division of labour between ‘static’ and dynamic approaches of the keylogs.

5. Potential Applications for Learning Corpus Research

One of the main interests of using keystroke logging to analyze research production is that it allows researchers to collect and analyze data live.

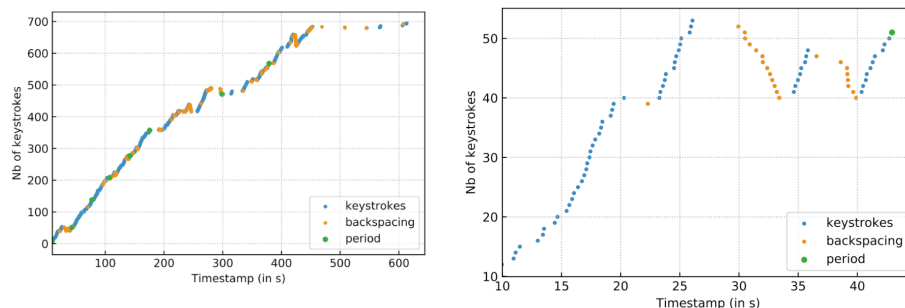


FIGURE 4. The potential complexity of backspacing and the need for a dynamic approach

This is particularly useful in an educational setting for it allows educators to provide visible feedback to students on the different aspects of their productions (Zhang, Zhu, Deane, and Guo, 2019).

Keystroke logging research presents interesting application possibilities, notably in language learning and teaching. When looking at the production of learners of a language in their target language, there are a few aspects that keystroke logging can inform us on that would not otherwise be accessible with only the final text as a resource. For instance, variables such as the amount of time spent on certain sections of the text or on difficult grammatical points are now available. It might also be interesting to look at how specific units such as reliability islands are produced. Editing, in the form of backspacing, is also made available by keystroke logging, which means that revision strategies are visible and can be analyzed.

When looking at keystroke data, it has been shown that four basic performance indicators were enough to separate typists into different clusters of learners that differed in writing processes and essay quality (*ibid.*). This could, in turn, lead teachers to better understand the needs of each specific student and to tailor their teaching to those needs.

Therefore, using keystroke features to investigate language production in an automated fashion will be useful to provide immediate and regular feedback to both students and educators.

This last section gives insight into learner data, perusing a portion of the data currently collected using Inputlog (Leijten and Van Waes, 2013) for the COREFL project (C. Lozano, A. Díaz-Negrillo, and Callies, to appear) at the university of Bremem to collect narratives. As can be seen in Fig. 5, writing bursts are not systematic.

In the second example (Fig. 6), we have manually represented the subdivisions of the writing task of a narrative based on a series of pic-

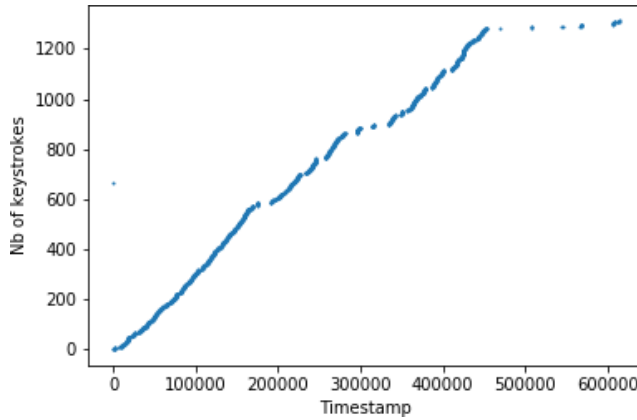


FIGURE 5. Visualization of typing bursts and picture changes in a narration task, data extracted from the COREFL corpus (A. M. C. Díaz-Negrillo and Cristóbal Lozano, 2018)

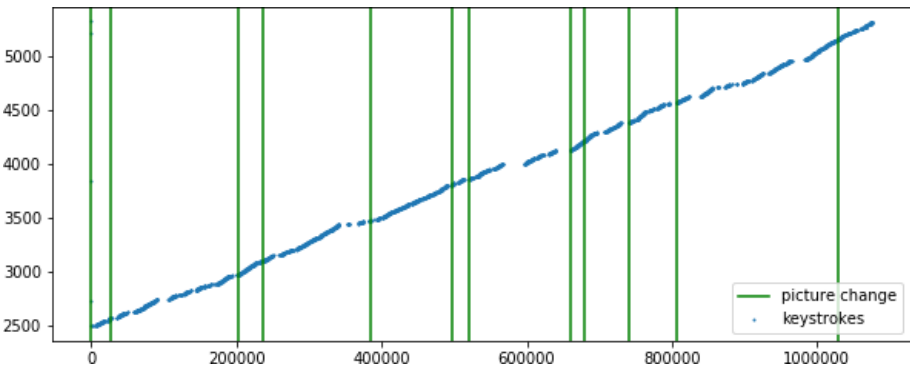


FIGURE 6. Visualization of typing bursts in time, data extracted from the COREFL corpus (A. M. C. Díaz-Negrillo and Cristóbal Lozano, 2018)

tures. In the figure, the vertical line represents a change from one picture to the other. As can be seen on the figure, some pictures require more description than others (as evidenced by the size of the window between each vertical green line), pauses may occur within one picture description, and the varying slopes correspond to different typing speeds for different pictures, which may in turn lead to question the difficulty of describing each picture.

6. Conclusion

In this chapter, we have suggested that the activity of writing with a keyboard shares features with speech in terms of potentially embedded constituents along a prosodic hierarchy. Our two-case studies with the two datasets considered allowed us to investigate only a fragment of the prosodic hierarchy. Whereas sublexical units such as suffixes have not seemed to be relevant, writing bursts and pauses call for investigations of units above the word such as collocations or reliability islands.

Analysing keystrokes gives an opportunity to reconsider Saussure's preference for speech over writing, as timepresses and time-release features act as features characterizing typed texts as time-stamped data, in a way similar to speech in spoken corpora. Aiming at analysing keylogs according to the prosodic hierarchy contextualises graphemes in relation to words, phrases, sentences and paragraphs, and therefore at text grammar level. It may not be the case that the variation of typing speed mirrors the variation of speech rhythm, but comparable grammars of chunking can be carried out for speech and keylog data.

References

- Bellis, Kouroch (2017). *La disposition Cœur 2.0 (ÉWOPY) comme disposition de clavier bureautique français: Réponse à l'enquête publique de l'AFNOR pour une norme PR NF Z71-300*. <https://hal.archives-ouvertes.fr/hal-01558613/document>.
- Bergadano, Francesco, Daniele Gunetti, and Claudia Picardi (2002). "User Authentication through Keystroke Dynamics". In: *ACM Transactions on Information and System Security (TISSEC)* 5.4, pp. 367–397.
- Charoenchaikorn, Vararin (2019). "L2 Revision and Post-task Anticipation during Text-Based Synchronous Computer-Mediated Communication (SCMC) Tasks". PhD Thesis. Lancaster University.
- Chukharev-Khudilaynen, Evgeny (2014). "Pauses in Spontaneous Written Communication: A Keystroke Logging Study". In: *Journal of Writing Research* 6.1, pp. 61–84.
- Cislaru, Georgeta and Thierry Olive (2016). "Les automatismes du scripteur: Jets textuels spontanés dans le processus de production écrite, le cas des constructions coordinatives". In: *SHS Web of Conferences*. Vol. 27. EDP Sciences, p. 06003.
- (2017). "Segments répétés, jets textuels et autres routines. Quel niveau de pré-construction?" In: *Corpus* 17, pp. 1–21.
- (2018). *Le processus de textualisation: analyse des unités linguistiques de performance écrite*. Louvain-la-Neuve, Paris: De Boeck Supérieur.

- Díaz-Negrillo, Ana Marcus Callies and Cristóbal Lozano (2018). "Designing and Compiling a Learner Corpus of Written and Spoken Narratives: The Corpus of English as a Foreign Language? (COREFL)". In: *ARISLA workshop (Anaphora Resolution in Second Language Acquisition)*. University of Granada.
- Evertz, Martin (in this volume). "The History of the Graphematic Foot in English and German".
- Giot, Romain, Mohamad El-Abed, and Christophe Rosenberger (2009). "Greyc Keystroke: A Benchmark for Keystroke Dynamics Biometric Systems". In: *2009 IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems (BTAS)*. Washington, DC, pp. 1–6.
- Giot, Romain et al. (2012). "Analysis of the Acquisition Process for Keystroke Dynamics". In: *Proceedings of the International Conference of the Biometrics Special Interest Group (BIOSIG)*. Darmstadt: IEEE, pp. 1–6.
- Kang, Pilsung and Sungzoon Cho (2015). "Keystroke Dynamics-Based User Authentication Using Long and Free Text Strings from Various Input Devices". In: *Information Sciences* 308, pp. 72–93.
- Killourhy, Kevin S. and Roy A. Moxon (2009). "Keystroke Dynamics—Benchmark Data Set". Carnegie-Mellon University, <http://www.cs.cmu.edu/~keystroke>.
- Leijten, Mariëlle and Luuk Van Waes (2013). "Keystroke Logging in Writing Research: Using Inputlog to Analyze and Visualize Writing Processes". In: *Written Communication* 30.3, pp. 358–392.
- Lozano, C., A. Díaz-Negrillo, and M. Callies (to appear). "Designing and Compiling a Learner Corpus of Written and Spoken Narratives: COREFL". In: *What's in a Narrative? Variation in Story-Telling at the Interface between Language and Literacy*. Ed. by Christiane Bongartz and Jacopo Torregrossa.
- Mahlow, Cerstin (2015). "Learning from Errors: Systematic Analysis of Complex Writing Errors for Improving Writing Technology". In: *Text, Speech and Language Technology*. Vol. 48: *Language Production, Cognition, and the Lexicon*. Springer, pp. 419–438.
- Malekian, Donia et al. (2019). "Characterising Students Writing Processes Using Temporal Keystroke Analysis". In: *The 12th International Conference on Educational Data Mining*. Ed. by Michel Desmarais et al. Vol. 27. Montréal, pp. 354–359.
- Monaco, John V. et al. (2013). "Behavioral Biometric Verification of Student Identity in Online Course Assessment and Authentication of Authors in Literary Works". In: *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*. Washington, DC, pp. 1–8.
- Nespor, Marina and Irene Vogel (2007). *Prosodic Phonology*. Vol. 28. de Gruyter.
- Plank, Barbara (2016). "Keystroke Dynamics as Signal for Shallow Syntactic Parsing". arXiv:1610.03321.

- Stewart, John C. et al. (2011). “An Investigation of Keystroke and Stylogometry Traits for Authenticating Online Test Takers”. In: *2011 International Joint Conference on Biometrics (IJCB)*. Washington, DC, pp. 1–7.
- Tappert, Charles C. et al. (2012). “A Keystroke Biometric System for Long-Text Input”. In: *Optimizing Information Security and Advancing Privacy Assurance: New Technologies*. Hershey, PA: IGI Global, pp. 32–57.
- Van Waes, Luuk, Mariëlle Leijten, and Christophe Neuwirth (2006). *Writing and Digital Media*. Leuven: Brill.
- Weingarten, Rüdiger, Guido Nottbusch, and Udo Will (2004). “Morphemes, Syllables, and Graphemes in Written Word Production”. In: *Trends in linguistics studies and monographs* 157, pp. 529–572.
- Zhang, Mo et al. (2019). “Identifying and Comparing Writing Process Patterns Using Keystroke Logs”. In: *Springer Proceedings in Mathematics & Statistics*. Vol. 265: *Quantitative Psychology*. Springer, pp. 367–381.