Pro-TEXT : an Annotated Corpus of Keystroke Logs

Aleksandra Miletic, Christophe Benzitoun, Georgeta Cislaru, Santiago Herrera-Yanez

Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1732–1739
Marseille, 20-25 June 2022

Pro-TEXT: an Annotated Corpus of Keystroke Logs

Aleksandra Miletić[1], Christophe Benzitoun[2], Georgeta Cislaru[3], Santiago Herrera-Yanez[1]
1 Clesthia (Paris 3 University), 2 ATILF (CNRS & Lorraine University),
3 MoDyCo (CNRS & Paris Nanterre University)
France
{aleksandra.haddad, georgeta.cislaru, santiago.herrera-yanez}@sorbonne-nouvelle.fr,
christophe.benzitoun@atilf.fr

Abstract

Pro-TEXT is a corpus of keystroke logs written in French. Keystroke logs are recordings of the writing process executed through a keyboard, which keep track of all actions taken by the writer (character additions, deletions, substitutions). As such, the Pro-TEXT corpus offers new insights into text genesis and underlying cognitive processes from the production perspective. A subset of the corpus is linguistically annotated with parts of speech, lemmas and syntactic dependencies, making it suitable for the study of interactions between linguistic and behavioural aspects of the writing process. The full corpus contains 202K tokens, while the annotated portion is currently 30K tokens large. The annotated content is progressively being made available in a database-like format, and the work on an HTML-based visualisation tool is currently under way. To the best of our knowledge, Pro-TEXT is the first corpus of its kind in French.

Keywords: annotated corpus, keystroke logs, annotation methodology, writing process, French

1. Introduction

This paper reports on annotation efforts on Pro-TEXT, a corpus based on keystroke logs written in French. Keystroke logs are recordings of the writing process executed on a keyboard and captured through dedicated software (Leijten and Van Waes, 2006; Strömqvist and Malmsten, 1998; Carl, 2012). These recordings keep track of all actions taken by the writer during the writing process (character additions and deletions, mouse movements, copy-paste substitutions, etc.), making them well-suited for data-based studies on the dynamics of the writing process and the underlying cognitive mechanisms. This is illustrated by the wide variety of research based on keystroke logs, ranging from studies on writing itself (Leijten and Van Waes, 2013; Ballier et al., 2018) to translation studies (Serbina et al., 2017; Carl et al., 2011) and language learning studies (Chukharev-Hudilainen et al., 2019; Miller et al., 2008).

Keystroke logs typically record rich behavioural information, such as pause duration between writing events and the speed of text sequence production, and Inputlog also provides some levels of linguistic annotation for English and Dutch (tokenization, lemmatization, POS-tagging, chunking and syllabification ; see (Leijten et al., 2015) for more details). For most other languages (including French), full linguistic annotation of keystroke logs needs to be done as a separate step. Given the often non-canonical nature of the data, which contains phenomena similar to disfluencies and error correction encountered in spoken language (Gilquin et al., 2011), automatic annotation poses similar challenges to those encountered when processing transcribed spoken corpora, e.g. (Gerdes and Kahane, 2009). It is therefore not surprising that fully annotated keystroke log corpora remain rare (see (Serbina et al., 2015) and (Carl, 2012) for two examples of POS-tagged keystroke logs).

Due to this lack of large amounts of annotated text, current studies of linguistic structures in keystroke logs are often based on manual inspection of smaller sets of data (e.g. (Cislaru and Olive, 2018)). Existing work on annotated data such as the one by Serbina et al. (2017) on word category changes in translation underline the importance of annotation.

Furthermore, data sharing does not seem to be a common practice for this type of corpora. This situation does not favour study comparability and reproducibility of results, nor does it foster the reuse of existing linguistic resources[1]. Our goals with the Pro-TEXT corpus are therefore as follows: create a rich database allowing further investigations into different aspects of the writing process, provide linguistic annotation for the corpus in order to make it suitable for quantitative, linguistically informed analyses, and make our data available for further research. To the best of our knowledge, it is the first such corpus for French.

The remainder of the paper is organized as follows: in Section 2, we give an overview of the corpus and the context in which it was created. In Section 3, we give a detailed account of the process on which we rely to annotate our data. In Section 4, we present the annotated part of the corpus and discuss some of its possible uses. We give our conclusions and directions for future work in Section 5.

[1] One of the exceptions is the CRITT-TPR database (Carl, 2012), which is publicly available.

2. Corpus Description

The Pro-TEXT corpus was built as part of the Pro-TEXT Project, an interdisciplinary project focusing on the writing process. The teams working on the project specialize in psycholinguistics (T. Olive, S. Bouriga, D. Chesnet, C. Perret, J. Pylouster and C. Bordes at CERCA, Poitiers University), linguistics (G. Cislaru, S. Fleury, F. Lefeuvre, D. Legallois, A. Boyer, Q. Feltgen and A. Miletic at CLESTHIA, Paris 3 University; C. Benzitoun and M. Dargnat at ATILF, Lorraine University), NLP (G. Cabanes, T. Charnois, N. Grozavu, J. Le Roux, P. Rastin, N. Rogovschi and N. Tomeh at LIPN, Paris 13 University) and translation studies (S. Vandaele, University of Montreal). Data collection was informed by the research orientation of the teams.

The corpus contains five subcorpora recorded in different conditions, with different types of authors. This diversity is intentional and serves the purpose of providing a source of information on different facets of the writing process.

Basic information on each subcorpus, including size, author profile and recording conditions, is available in Table 1, and more details are provided below. The word counts given in the table refer to the final versions of the texts.

The data was recorded in real time using two keystroke logging programs: InputLog (Leijten and Van Waes, 2006)[2], which runs on Windows, and Scriptlog (Strömqvist and Malmsten, 1998)[3], which runs on macOS.

The degree of writing expertise was established based on the duration in years of the daily practice of writing and on the expected degree of proficiency with respect to the discourse genres that were to be produced by the author. For instance, all adults who produced non-specialized texts on general subjects were considered experts. Students who were asked to produce mini-research papers were considered semi-experts, given that they were proficient in writing as a general practice, but had not yet mastered the specific task of writing academic texts. We did not assess language or writing skills before the data recording process.

2.1. Subcorpus *Academic*

This subcorpus contains mini-theses written by MA students as part of a course in discourse analysis. The texts were written over several writing sessions, on students' computers. Since this type of writing task was novel to the participants, they were evaluated as semi-experts. The students involved in data collection were native or near-native speakers of French. There are 26 different authors in the subcorpus.

2.2. Subcorpus *Professional*

Reports on child protection were written by social workers as part of their regular tasks. The reports were written over several sessions, and one text can have several authors. Since the participants wrote these types of texts routinely, they were evaluated as experts. There are 9 different authors, and they were all native speakers of French.

2.3. Subcorpus *Experimental*

These texts were produced as part of a psycholinguistic experiment on the writing process. They were written by BA students. The texts are essays on different social topics, such as smoking at the university and public transportation. There were three experimental conditions, focused respectively on the stages of planning, producing and revising the text. In each condition, each author produced one text in experimental conditions and one in control conditions. Each text was written in a single session. Since this type of writing task is common in the French educational system, the authors were evaluated as experts. The information about the experimental setting and experimental vs control setting is available for each text. There are 83 authors in this subcorpus, and they are all native or near-native speakers of French.

2.4. Subcorpus *Children*

The texts in this part of the corpus were written by schoolchildren from three age groups: 3rd year of primary school (ca. 8 years old), (5th year of primary school (ca. 10 years old), and 1st year of secondary school (ca. 11 years old). Each participant wrote a narrative text and an essay on a given subject. The texts were recorded at school, in one writing session. The information about the age group, the type of text and the order of the production of the two texts is available for each text. There are 92 authors in total, and they are considered to hold a language proficiency level corresponding to their grade.

2.5. Subcorpus *Translation*

This subcorpus was written by BA students of translation studies. Each participant produced two types of text: an original text in French describing an image, and a translation of a medical text from English to French. The information on the author and on the type of text is available for each text in the subcorpus. Given the type of the task and the fact that the text had to be produced in a highly specialized discourse genre, the students were evaluated as semi-experts. There are 19 authors in total in this subcorpus and they have native or near-native proficiency level in French.

A part of this content was selected for the annotation process, which is described in the remainder of this paper.

3. Annotation Methodology

As mentioned in Section 1, keystroke log corpora seem to be rarely annotated, and the existing annotations are

[2] Available at `http://www.inputlog.net/`

[3] Available upon request.

Subcorpus	Texts	Words	Writers	Genre	Expertise
Academic	26	70464	MA students	mini-thesis in linguistics	semi-experts
Professional	10	34504	social workers	reports on child protection	experts
Experimental	165	63533	BA students	essays on different subjects	experts
Children	183	20306	pupils (3rd - 6th grade)	narrative texts and essays	beginners
Translation	38	13682	BA students	EN-FR translation of medical texts and original texts produced in FR	semi-experts
Total	422	202489	-	-	-

Table 1: Content of the Pro-TEXT corpus

almost exclusively done on the final text (Carl, 2012). However, the added value of this type of corpora resides precisely in the fact that they also record the dynamics of the writing process, captured as intermediate versions of texts: all of the modifications made by the writer during the writing process are available. In other words, a sentence in the final text may correspond to several intermediate versions captured in the log data, such as in Example 1. Here, each subexample corresponds to a successive intermediate version of the same sentence. The deletions are marked with strikethrough font, and additions with respect to the preivous intermediate version are given in bold. The modifications between versions can be as diverse as replacing a constituent (cf. 1a vs 1b vs 1c), correcting spelling (cf. 1c vs 1d), or modifying a lexical choice (cf. 1e vs 1f vs 1g).

(1) a. afin d' aborder un projet de l'
in order to address a project of the
Université de Poitiers
University of Poitiers

b. afin d' aborder un projet ~~de l'~~
in order to address a project ~~of~~ ~~the~~
~~Université~~ ~~de Poitiers~~
~~University~~ ~~of~~ ~~Poitiers~~

c. afin d' aborder un projet **songé par**
in order to address a project **conceived by**
l' Universitté
the Universitty

d. afin d' aborder un projet songé par
in order to address a project conceived by
l' Universitté
the Universitty

e. afin d' aborder un projet songé par
in order to address a project conceived by
l' Université **de Poitiers**
the University **of Poitiers**

f. afin d' aborder un projet songé par
in order to address a project conceived by
~~l'~~ ~~Université~~ ~~de Poitiers~~
~~the~~ ~~University~~ ~~of~~ ~~Poitiers~~

g. afin d' aborder un projet songé par
in order to address a project conceived by
notre Université
our University

In order to maximize the potential of keystroke log corpora for linguistic research, it is essential to also annotate the parts of the content that do not make it into the final version. This need was taken into account e.g. by Serbina et al. (2015) in their work on a keystroke log corpus of translations. Our first goal is therefore to annotate all content produced by the writer and not only the final text.
Second, in order to make the corpus as reliable a source of information as possible, we check and validate the annotation manually. To mitigate the fact that such an approach is highly time-consuming, we combine two annotation strategies: automatic data pre-annotation (to accelerate manual annotation) and agile annotation (to ensure manual annotation quality).

3.1. Annotating All of the Content: Final Texts and Intermediate Versions

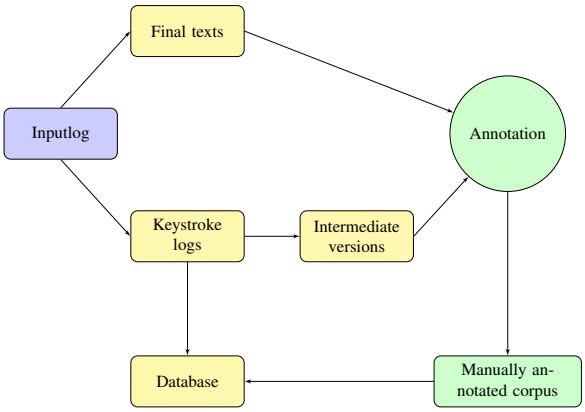

Figure 1: Global overview of the annotation process

A global overview of our annotation process is given in Figure 1. Inputlog generates two main types of output: a file with the final version of the produced text, and a corresponding keystroke log file[4]. The keystroke logs are transformed into a database containing behavioural

[4]If a text is produced through several writing sessions, there will be a log file per session.

information such as writing speed, pause length and timestamps for each writing event. The database is currently organized on character level: each entry corresponds to a character (or deletion action) produced by the writer. Using the database, it is possible to locate each writing event in the spatial dimension of the text, which is crucial for the HTML visualization. The work on extracting the database from the keystroke logs is done by C. Bordes (Poitiers University).

In the annotation process, our goal is to enrich this database with more detailed linguistic annotation. To do so, we adopt a two-step approach. We first annotate the final version of each text. Since these texts are in principle regular, they are relatively easy both for automatic preprocessing and for manual validation. In the second step, we first use the keystroke logs to reconstitute intermediate versions of each text. We then generate an automatic annotation for each intermediate version.

Serbina et al. (2015) noted that intermediate versions pose a particular challenge for automatic annotation: the input for the annotation process includes partially written tokens, incomplete syntactic structures, erroneous wordforms to be corrected (or not) later on in the writing process. This type of data is therefore comparable to non canonical linguistic material such as computer-mediated communication or learners' writing, which is notoriously difficult for NLP tools. We can add that these data also contain some phenomena comparable to disfluencies and repairs in spoken corpora, which are also challenging for current annotation tools (Gerdes and Kahane, 2009; Kahane and Gerdes, 2020).

To mitigate this effect, Serbina et al. (2015) perform manual annotation checks. We adopt the same approach but diverge from the methodology proposed by Serbina et al. (2015) on one important point. They chose to generate an annotation for a token after each text-modifying event (addition or deletion of a character). However, this produces an important amount of additional data (potentially, a new POS tag is added to the annotation of a word with each typed character). In order to simplify this process and reduce the amount of data, we produce a new annotation after each series of text-modifying events of the same type occurring on adjacent positions in the text. In other words, we re-annotate a sentence after each series of insertions or deletions at the same point in the existing material. The intermediate versions in Example 1 follow these rules.

Since an important part of intermediate versions (sentences and parts of sentences) also appear in the final text, we use this fact to project the manually corrected annotation of the final text onto the intermediate versions. The sequences that **do not** appear in the final text are left with the automatic annotation and need to be corrected manually. This represents only a fraction of the complete content of intermediate versions.

In the final step, the full annotation is projected back onto the initial database. Thus, the behavioural information and the linguistic annotation can be used together. A CSV-based file intended for quantitative analysis and machine learning experiments, as well as a CoNLL file containing the annotation, are created for each text. The work on an HTML-based visualization is under way; the display of the data will be derived from the CSV database.

3.2. Making It Easier for Annotators: Automatic Pre-annotation

In order to facilitate the task for human annotators, we rely on automatic pre-annotation of our data. This approach is supported by the well-established positive effects the method has on various types of linguistic annotation (Xue et al., 2005; Fort, 2012; Tellier et al., 2014; Miletic et al., 2019; Miletić et al., 2020). Pre-processing is done with the Talismane NLP pipeline (Urieli, 2013). Although more recent tools are available, we chose Talismane for several reasons. First, it had already been used on French with solid results (Urieli, 2013). Also, it works as a full processing pipeline, able to transform running text into fully annotated dependency trees. Finally, a Talisman model trained on the tagsets we wanted to use was already available, making the annotation setup quicker. We use the models distributed with the tool, trained on the French Treebank (Candito et al., 2009). Therefore, the POS-tagset and the dependency label set are the ones used in that corpus. An overview of the tags and labels we use is given in the Appendix (Tables 4 and 5, respectively).

Several annotation layers are generated with the tool: sentence segmentation, tokenization, POS-tagging, lemmatization, and dependency parsing. The dependency annotation is filtered based on the probability score assigned by the tool in order to minimize the noise in the pre-annotation layer. The annotators then manually correct and complete the annotation through the Arborator-Grew interface (Guibon et al., 2020).

3.3. Ensuring Annotation Quality: Agile Annotation

A more detailed representation of the annotation organization is given in Figure 2. Following Fort (2012), we divide the annotation work into four stages: campaign preparation (blue), pre-campaign (yellow), manual validation campaign (green) and corpus finalisation (red). For the campaign stage of the process, we adopt the agile annotation approach defined by Voormann and Gut (2008): annotation is iterative, with each iteration followed by an evaluation step, the role of which is to ensure the quality of the produced annotation.

1. **Campaign preparation** included selecting texts to be annotated, choosing pre-annotation tools and the manual validation interface, and preparing the initial version of the annotation guidelines.

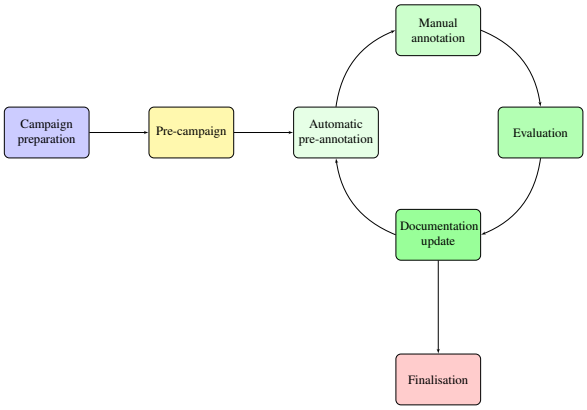

Figure 2: Annotation campaign organization

2. **Pre-campaign** involved recruiting annotators and training them on the guidelines and the use of the annotation interface. Since we use automatic pre-annotation, pre-campaign also included automatic data pre-processing.

3. **Annotation campaign** comprised iterative cycles of manual annotation and evaluation. The evaluation step consisted in organizing regular annotator meetings dedicated to resolving problematic cases and validating annotation decisions. Annotation guidelines were regularly updated based on these discussions.

4. **Finalisation** involves final annotation coherency checks and leads to corpus distribution. As the annotation guidelines were updated after each annotation cycle, it is essential to harmonize annotations in order to ensure coherent linguistic analysis throughout the corpus. Once this step is done, the validated part of the corpus is published.

Voormann and Gut (2008) recommend calculating inter-annotator agreement as part of each evaluation step. In our case, this was not done during the annotation of final texts because the annotation guidelines were still evolving. We relied instead on annotator meetings to ensure annotation quality. However, calculating inter-annotator agreement will be included in the campaigns dedicated to the manual correction of intermediate versions.

4. Annotated Corpus and Project Status

At the moment of writing, the final version of 147 texts containing 30146 words has been automatically annotated and manually validated. Details on the distribution of annotated texts across subcorpora and some basic statistics are available in Table 2.
Note that the unexpectedly high mean sentence length in the subcorpus *Children* is probably due to the unsystematic use of punctuation among young writers. This often results in texts that are a single graphical sentence.

The absence of texts from the subcorpus *Professional* is due to anonymisation issues. As mentioned in Section 2, these files are reports about social protection of children. As such, they contain highly sensitive information about individuals, and we are currently ensuring that each intermediary version is fully anonymized before being processed and published. A sample of this subcorpus will be available at the end of the annotation process.

Some further information on the distribution of POS tags and dependency labels across this part of the corpus is available in the Appendix (cf. Tables 4 and 5).
The second annotation step, in which the intermediate versions of texts are annotated, is under way. Currently, 49 texts from the *Children* subcorpus have been processed automatically and the manual correction of deleted sequences is ongoing. Basic information about this sample is available in Table 3, both for the final texts and for the intermediate versions. Note that the intermediate versions also contain the final version of the given text. Each final text is therefore a subset of the tokens and annotations available in the corresponding intermediate version file.
The corpus in its current state is available under the Creative Commons BY-NC-SA 4.0 licence. It can be downloaded from the site of the project: `https://pro-text.huma-num.fr/ressources/`.
In the following months, the annotation efforts will be focused on producing the full annotation of intermediate versions for the final texts that have already been manually validated. The remainder of the corpus will be annotated according to the same methodology. We are currently exploring the possibility of using the annotated data to leverage a bootstrapping approach in a method comparable to (Kahane and Gerdes, 2020) in the hope of improving the quality of the automatic annotation.

5. Conclusions and Future Work

In this paper, we presented the methodology for enriching Pro-TEXT, a keystroke log based corpus written in French, with linguistic annotation. This methodology allows us to enrich behavioural information recorded in keystroke logs with several layers of linguistic information (lemmas, POS-tags, syntactic dependencies). Special attention was given to reducing the amount of manual work required from human annotators and to ensuring annotation quality.
The creation of this corpus opens promising avenues for new research. Among many other possibilities, annotating the data will allow us to examine and describe the nature of writing bursts, to observe the correlations between the linguistic structure and the segmentation of the production flow by pauses, and to examine the behaviour of syntactic dependencies with respect to the writing dynamics. The interactions between behavioural data and linguistic annotation will also be modelled using machine learning techniques.

Subcorpus	Texts	Sentences	Tokens	Token/sentence	Lemmas	Types
Children	120	440	11873	27.0	1171	224
Experimental	15	149	5719	38.4	1024	1428
Translation	10	149	3675	24.7	626	887
Academic	2	474	8879	18.7	1523	2013
Professional	0	0	0	0	0	0
TOTAL	147	1212	30146	24.9	3062	5184

Table 2: Annotated Final Texts Statistics

	Sentences	Tokens	Lemmas	Types
Final texts	71	4319	632	1074
Intermediate versions	4621	128518	693	1767

Table 3: Annotated Intermediate Version Statistics

The annotation of the full corpus is ongoing. The currently annotated content is available for download. We hope that sharing our data will help foster resource reuse and result comparability in the domain of writing research.

6. Acknowledgements

This work was funded through the ANR project *Processes of Textualization: Linguistic, Psycholinguistic, and Machine Learning Modeling* (project N° ANR-18-CE23-0024-01). More details are available at: `https://pro-text.huma-num.fr/le-projet/`.

7. Bibliographical References

Ballier, N., Pacquetet, E., and Arnold, T. (2018). Investigating keylogs as time-stamped graphemics. In *Graphemics in the 21st Century 2018*, pages 353–365.

Carl, M., Dragsted, B., Elming, J., Hardt, D., and Jakobsen, A. L. (2011). The process of post-editing: A pilot study. *Copenhagen Studies in Language*, 41:131–142.

Carl, M. (2012). Translog-ii: a program for recording user activity data for empirical reading and writing research. In *LREC*, volume 12, pages 4108–4112.

Chukharev-Hudilainen, E., Saricaoglu, A., Torrance, M., and Feng, H.-H. (2019). Combined deployable keystroke logging and eyetracking for investigating l2 writing fluency. *Studies in Second Language Acquisition*, 41(3):583–604.

Cislaru, G. and Olive, T. (2018). *Le processus de textualisation: analyse des unités linguistiques de performance écrite*. De Boeck Supérieur.

Fort, K. (2012). *Les ressources annotées, un enjeu pour l'analyse de contenu: vers une méthodologie de l'annotation manuelle de corpus*. Ph.D. thesis, Université Paris-Nord-Paris XIII.

Gerdes, K. and Kahane, S. (2009). Speaking in piles: Paradigmatic annotation of french spoken corpus. In *Fifth Corpus Linguistics Conference*, pages 1–15.

Gilquin, G., De Cock, S., et al. (2011). Errors and disfluencies in spoken corpora: Setting the scene. *International Journal of Corpus Linguistics*, 16(2):141–172.

Guibon, G., Courtin, M., Gerdes, K., and Guillaume, B. (2020). When collaborative treebank curation meets graph grammars. In *LREC 2020-12th Language Resources and Evaluation Conference*.

Kahane, S. and Gerdes, K. (2020). Annotation syntaxique du français parlé: Les choix d'orféo. *Langages*, (3):69–86.

Leijten, M. and Van Waes, L. (2006). Inputlog: New perspectives on the logging of on-line writing processes in a windows environment. In *Computer keystroke logging and writing*, pages 73–93. Brill.

Leijten, M. and Van Waes, L. (2013). Keystroke logging in writing research: Using inputlog to analyze and visualize writing processes. *Written Communication*, 30(3):358–392.

Leijten, M., Van Waes, L., and Van Horenbeeck, E., (2015). *Writing(s) at the Crossroads: The Process-Product Interface*, chapter Analyzing writing process data: A linguistic perspective, pages 277–302. John Benjamins Publishing Company.

Miletic, A., Fabre, C., and Stosic, D. (2019). De la constitution d'un corpus arboré à l'analyse syntaxique du serbe. *Traitement Automatique des Langues*, January.

Miletić, A., Bras, M., Vergez-Couret, M., Esher, L., Poujade, C., and Sibille, J. (2020). Building a universal dependencies treebank for occitan. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2932–2939.

Miller, K. S., Lindgren, E., and Sullivan, K. P. (2008). The psycholinguistic dimension in second language writing: Opportunities for research and pedagogy using computer keystroke logging. *Tesol Quarterly*, 42(3):433–454.

Serbina, T., Niemietz, P., Fricke, M., Meisen, P., and Neumann, S. (2015). Part of speech annotation of

intermediate versions in the keystroke logged translation corpus. In *Proceedings of the 9th LinguiLtic Annotation Workshop*, pages 102–111.

Serbina, T., Hintzen, S., Niemietz, P., and Neumann, S., (2017). *Empirical modelling of translation and interpreting*, volume 3, chapter Changes of word class during translation–Insights from a combined analysis of corpus, keystroke logging and eye-tracking data, pages 177–208. Language Science Press, Berlin.

Strömqvist, S. and Malmsten, L. (1998). Sriptlog pro 1.04: User's manual. Technical report, University of Göteborg.

Tellier, I., Eshkol-Taravella, I., Dupont, Y., and Wang, I. (2014). Peut-on bien chunker avec de mauvaises étiquettes POS? In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN2014)*, pages 125–136, Marseilles, France. Association pour le Traitement Automatique des Langues (ATALA).

Urieli, A. (2013). *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Ph.D. thesis, Université Toulouse le Mirail-Toulouse II.

Voormann, H. and Gut, U. (2008). Agile corpus creation. *Corpus Linguistics and Linguistic Theory*, 4(2):235–251.

Xue, N., Xia, F., Chiou, F.-D., and Palmer, M. (2005). The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. *Natural language engineering*, 11(02):207–238.

8. Language Resource References

Candito, Marie and Crabbé, Benoît and Falco, Mathieu. (2009). *Dépendances syntaxiques de surface pour le français*.

Carl, Michael. (2012). *The CRITT TPR-DB 1.0: A database for empirical human translation process research*.

Appendix

POS tag	Meaning	Count
ADJ	non-interrogative, non-relative adjective	1358
ADJ—VPP	ambiguous form that can be a past participle or an adjective	39
ADV	non-interrogative adverb	2001
ADVWH	interrogative adverb	30
CC	coordinating conjunction	985
CLO	object clitic	562
CLR	reflexive clitic	320
CLS	subject clitic	1502
CS	subordinating conjunction	682
DET	non-interrogative determiner	3815
DETWH	interrogative determiner	4
ET	foreign language content	31
I	interjection	24
NC	common noun	5229
NPP	proper noun	549
NUM	numeral	153
P	preposition	3189
P+D	preposition+determiner	441
P+PRO	preposition+pronoun	5
PONCT	punctuation	3123
PRO	non-interrogative, non-relative pronoun	663
PROREL	relative pronoun	385
PROWH	interrogative pronoun	37
V	indicative verb	2932
VIMP	imperative verb	35
VINF	infinitive verb	1023
VPP	past participle	710
VPR	present participle	77
VS	subjunctive verb	43

Table 4: POS-tags in the annotated final texts

Dep. label	Meaning	Count
a_obj	indirect object introduced by *à*	353
aff	affix	309
ap	apposition	173
arg	argument of a fixed prepositional construction	3
arg_comp	argument of a comparative construction	15
ato	direct object complement	25
ats	subject complement	508
aux_caus	causative auxiliary	52
aux_pass	passive auxiliary	152
aux_tps	temporal auxiliary	482
comp	completive subordinate clause	237
coord	coordinating conjunction	656
de_obj	indirect object introduced by *de*	126
dep	prepositional dependent of a noun	1393
dep_coord	conjunct in a coordination	1163
det	determiner	3804
detachement	complement in a detached construction	12
fixed	element of a multiword expression	393
goeswith	character sequence that belongs to an immediately preceding word	134
mod	modifier (of a verb or a noun)	4851
mod_cleft	cleft clause	32
mod_rel	relative clause	354
obj	direct object	2357
p_obj	prepositional indirect object	356
prep	preposition	3458
root	sentence root	1889
sub	adverbial subordinate clause	409
suj	subject	2691
suj_impers	subject in an impersonal construction	298
unknown	syntactic function impossible to determine	33

Table 5: Dependency labels in the annotated final texts