Pro-TEXT Project
Aleksandra Miletic
14/04/2022

ProTEXT Project: annotated corpus

1. Project scope and goals

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.

2. Corpus annotation

Keystroke logs are recordings of the writing process executed on a keyboard and captured through dedicated software. They typically record rich behavioural information, such as pause duration between writing events and the speed of text sequence production. Combining this type of data with linguistic annotation would make them an excellent source for data-based studies on the dynamics of the writing process, linguistic structures involved in it and underlying cognitive mechanisms.
However, keystroke log corpora seem to be rarely annotated, and the existing annotations are almost exclusively done on the final text (Carl, 2012). The added value of this type of corpora resides precisely in the fact that they also record the intermediate versions of texts: all of the modifications made by the writer during the writing process are recorded. In other words, a sentence in the final text may correspond to several intermediate versions captured in the log data, such as in Figure 1. Here, each subexample corresponds to successive intermediate versions of the same sentence.

(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 songé par
in order to address a project thought by
l' Universitté
the Université

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

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

Figure 1: Intermediary versions of a sentence

In order to maximize the potential of keystroke log corpora for linguistic research, it is essential to also annotate the parts of content that do not make it into the final version. Our main goal is therefore to annotate all content produced by the writer, and not only the final text.

2.1 Annotation methodology

The Pro-TEXT corpus was collected using Inputlog (Leijten and Van Waes, 2006). This software provides two types of output: text files containing the final version of the text and IDFX files (an XML format) containing keystroke logs.
In order to annotate both the final text and the text sequences that were deleted during the writing process, we adopt a two-step approach, given in Figure 2.

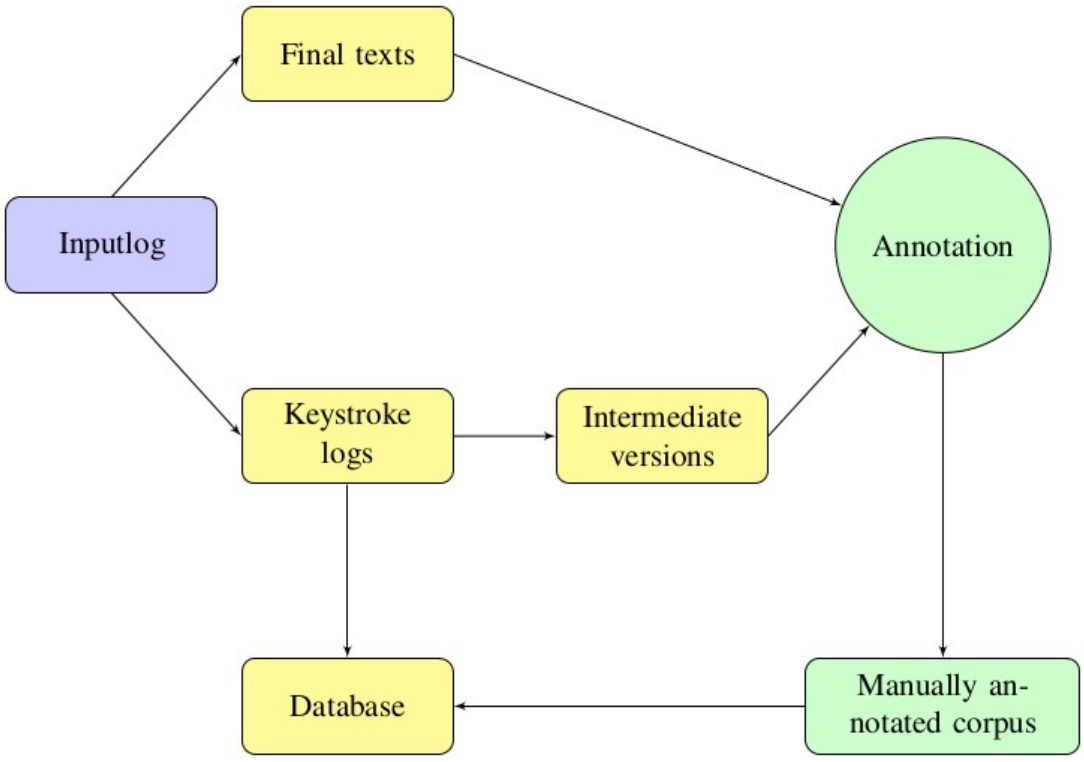

Figure 2: Annotation process

1. We first annotate the final version of each text. These are pre-annotated automatically and corrected manually. The result is stored in a CoNLL format (column-based text format).
2. In the second step, we transform the keystroke logs into explicit intermediary versions of each text. These are then annotated automatically. Note that an important part of intermediary versions (sentences and parts of sentences) also appear in the final text. We therefore project the manually corrected annotation of the final text onto the intermediary versions. Currently, this leaves the sequences that DO NOT appear in the final version (i.e. that were deleted or substituted later on)

with automatic annotation only. The manual correction for these parts is under way and corrected files will be distributed as soon as they are available. The resulting output files are detailed in section 3.

2.2 Annotation tagsets

The corpus is annotated with POS tags, lemmas and syntactic dependencies. The tables below give a brief overview of the tagsets used for POS tagging and syntactic annotation.

POS tag	Meaning
ADJ	non-interrogative, non-relative adjective
ADJ—VPP	ambiguous form that can be a past participle or an adjective
ADV	non-interrogative adverb
ADVWH	interrogative adverb
CC	coordinating conjunction
CLO	object clitic
CLR	reflexive clitic
CLS	subject clitic
CS	subordinating conjunction
DET	non-interrogative determiner
DETWH	interrogative determiner
ET	foreign language content
I	interjection
NC	common noun
NPP	proper noun
NUM	numeral
P	preposition
P+D	preposition+determiner
P+PRO	preposition+pronoun
PONCT	punctuation
PRO	non-interrogative, non-relative pronoun
PROREL	relative pronoun
PROWH	interrogative pronoun
V	indicative verb
VIMP	imperative verb
VINF	infinitive verb
VPP	past participle
VPR	present participle
VS	subjunctive verb

Dep. label	Meaning
a_obj	indirect object introduced by *à*
aff	affix
ap	apposition
arg	argument of a fixed prepositional construction
arg$_c omp$	argument of a comparative construction
ato	direct object complement
ats	subject complement
aux$_c aus$	causative auxiliary
aux$_p ass$	passive auxiliary
aux$_t ps$	temporal auxiliary
comp	completive subordinate clause
coord	coordinating conjunction
de_obj	indirect object introduced by *de*
dep	prepositional dependent of a noun
dep_coord	conjunct in a coordination
det	determiner
detachement	complement in a detached construction
fixed	element of a multiword expression
goeswith	character sequence that belongs to an immediately preceding word
mod	modifier (of a verb or a noun)
mod_cleft	cleft clause
mod_rel	relative clause
obj	direct object
p_obj	prepositional indirect object
prep	preposition
root	sentence root
sub	adverbial subordinate clause
suj	subject
suj_impers	subject in an impersonal construction
unknown	syntactic function impossible to determine

3. File formats

The annotation process described above yields two main types of output: a CoNLL file containing the annotation of all intermediary versions, and a CSV file containing a character-based database that combines the annotation with the behavioural information derived from keystroke logs.

3.1 Annotated intermediary versions in CoNLL format

The CoNLL format is a column-based text format in which each row contains one token (or word), and different columns contain different levels of annotation. The meaning of the columns is the following:

1. **token ID**: the ID of the token; in the Pro-TEXT corpus, it is unique on text-level. The uniqueness of the token is established through the IDs of the characters that compose it (see below).
2. **token**: the word itself
3. **lemma**: the dictionary form of the word
4. **POS-tag**: grammatical category
5. **XPOS-tag**: an alternative grammatical tag, typically used in converted corpora in order to keep the old annotation. Here, it typically contains the non-corrected POS-tag.
6. **Morphosyntactic features**: automatically produced morphosyntactic traits such as number, gender or person. **These are not corrected manually at any stage.**
7. **governor**: the syntactic governor of the word, identified through its ID.
8. **function**: the syntactic function of the word with respect to its governor
9. this column typically contains **additional dependency relations** (often the deep syntactic relations). In our case, it is typically empty and used only to indicate a potentially incomplete token in an intermediary version. If a token is considered as such, the column contains the indication '***'. This is used to facilitate the identification of such tokens during the manual correction step.
10. **miscellaneous**: this column allows to add other types of information. In the Pro-TEXT corpus, it contains the information on the characters that constitute the token. The format is as follows: [text-level unique character ID]=[character]__[status in final text]. The character ID allows to link each character to the database with behavioural information. The presence of the character itself allows for easy checks of character ID projections. The character status refers to the presence (True) or absence (False) of the character in the final version of the text. Any token containing a character with status *False* does not belong to the final text.

The beginning of each intermediary version is signaled by the metadata line starting with '#text_version='. The number -1 indicates the final version of the text.

Each new sentence is signaled by the metadata line starting with '#sentence_id='. Note that the final version in the children's corpus was not segmented into sentences, given the highly irregular use of punctuation marks in this subcorpus.

These files can be visualized through an annotation interface such as Arborator (https://arboratorgrew.elizia.net/#/) or queried through appropriate tools. However, they do not integrate the behavioural data.

Pro-TEXT Project
Aleksandra Miletic
14/04/2022

3.2 Character-based databases in CSV format

The character-based databases in CSV format combine the behavioural information from keystroke logs with the linguistic annotation. Each text-modifying action (addition or deletion) from the IDFX file is assigned a unique character ID, which allows to link it to CoNLL files. The columns contain behavioural information and annotations for a given character for each intermediary version.

Behavioural information columns:

1. **ID**: text ID
2. **session**: writing session (currently, only single-session texts are annotated)
3. **writer**: writer ID
4. **n_event**: the number of the event in the IDFX log. Keyboard events carry distinct event numbers. Actions carried out through select-delete or copy-paste can concern several characters, but they all have the same event number.
5. **st_time**: start time of the event
6. **end_time**: end time of the event
7. **pause**: time between the end of the previous event and the start of the current event
8. **event**: event itself (the character produced or the symbol indicating deletion)
9. **pos**: position in text at which the event occurred.
10. **doc_len**: document length in characters **before** the current event
11. **op**: opearation (1=addition, -1=deletion, 0=does not change the length of the document)
12. **type_op**: keyboard or insert/replace

Annotation columns:

1. **charID**: text-level unique character ID (attributed in the order of the production)
2. **tokenID**: text-level unique token ID of the token to which the character belongs
3. **token**: token to which the current character belongs
4. **lemma**: lemma of the token to which the character belongs
5. **POS**: POS tag of the token to which the character belongs
6. **XPOS**: non corrected tag of the token to which the character belongs
7. **ms**: morphosyntactic features of the token to which the character belongs
8. **governor**: syntactic governor of the token to which the character belongs
9. **function**: syntactic function of the token to which the character belongs
10. **charStatus**: status of the character in the final text (True=present, False=absent)
11. **tokenStatus**: indicates if the token is considered incomplete (***) or not (_)
12. **sentence_id**: ID of the sentence to which the character belongs

All annotation columns except for charID and charStatus give the information for each intermediary version of the text. This is done using the following format:
[text version]=[annotation](|[text version2]=[annotation])*

4. Available files

This section gives corpus statistics for currently available files. In version 0.1, only texts from the *children* subcorpus are available.

4.1 *Children* subcorpus

This subcorpus contains texts written by school children. There three age groups: 3rd year of primary school (8 years old), 5th year of primary school (10 years old) and 1st year of secondary school (11 years old). Participants produced two types of texts: an argumentative essay and a narrative text. These pieces of information are encoded in the text titles, which follow the format *[participantID][age group][type of text][order of production]*.

Participant ID:
P[0-9]+: unique ID of the participant inside their age group
Age group:
CE2: 3rd year of primary school (8 years old)
CM2: 5th year of primary school (10 years old)
C6: 1st year of secondary school (11 years old)
Type of text:
N = narrative
E = argumentative
Order of production:
1 = the first text to be produced by the participant
2= the second text to be produced by the participant.

A total of 49 annotated texts are currently available. Basic corpus statistics are given in the table below, both for the final texts and for the intermediary versions. Note that the files with the intermediary versions also contain the final version of the text. The final texts are therefore a subset of the tokens and annotations available in the intermediary version files.

	tokens	lemmas	token/sentence
final texts	4319	632	60.8
intermediary versions	128518	693	27.8

Table 1: Corpus statistics for annotated files from the children subcorpus

5. References

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

Carl, M. (2012). The CRITT TPR-DB 1.0: A database for empirical human translation process research. Workshop on Post-Editing Technology and Practice