

In search of constructions in writing process data

Gaëtanelle Gilquin, Université catholique de Louvain

Abstract

This article explores the possibility of identifying mentally stored constructions in writing process data, that is, data that reproduce the process through which a text is written. The unit that serves as a basis for the identification of constructions is the burst of writing, which corresponds to a chunk of text produced between two pauses. Bursts are examined in L1 French and L2 English keylogging data from the Process Corpus of English in Education and their potential constructional status is considered.

Keywords: Construction Grammar, writing process, keylogging, burst of writing

1. Introduction

Constructions, in the (Cognitive) Construction Grammar sense, are linguistic patterns that correspond to form-meaning pairings stored in the mind or, more precisely, in the mental constructicon (see Goldberg 2003, 219). Provided there is agreement on the theoretical definition of constructions, deciding what linguistic items represent a construction does not seem to be much of an issue among construction grammarians. This is surprising because, while the existence of a form-meaning pairing can be

established fairly easily by looking at language, determining whether this pairing is mentally stored or not involves some kind of access to the mind – not an easy task given that the human mind is pretty much a black box.

This article surveys the main types of approaches that have been adopted to identify constructions and then examines the possibility of identifying constructions in writing process data, that is, data reflecting the process through which a text is composed. The unit that serves as a basis for construction identification in these data is the ‘burst’ (see Chenoweth and Hayes 2001), that is, a chunk of text produced between two pauses. Keylogging data from the Process Corpus of English in Education (PROCEED) help explore this idea.

2. Identification of constructions

Many constructions – and this was especially the case in the early days of Construction Grammar (CxG) – are identified on the basis of linguists’ introspection, often combined with linguistic principles.¹ The famous example “Pat sneezed the napkin off the table” (Goldberg 1995, 3), for instance, is taken as evidence that there is a so-called caused motion construction, since the verb *sneeze*, which is normally intransitive, can be used in this pattern while still making sense. An important element in the definition and hence identification of constructions is their non-compositionality, that is, the fact that the form and/or meaning of a construction is not fully predictable from its components. Some studies have therefore used non-compositionality as an argument to claim that a linguistic phenomenon is indeed a construction. Hilpert and Bourgeois (2020, 99),

talking about the *sarcastic much?* pattern (e.g. ‘Angry much?’), point out that “[t]he idiosyncratic and non-compositional characteristics that make this pattern a construction (...) are in plain sight”. While this type of reasoning follows the theoretical claims of CxG, the compositionality or non-compositionality of a pattern does not in itself prove whether the pattern is stored in the mind. As rightly emphasised by Croft (1998, 151), claims about mental representation cannot be made “without evidence from actual usage or psycholinguistic experiments”.

Both usage and experimentation have been used in CxG studies to identify constructions empirically. The former is reflected in corpora, which make it possible, among others, to measure the frequency of linguistic patterns. According to the tenets of CxG, “sufficient frequency” (Goldberg 2006, 5) should be a sign that a pattern (even a fully compositional one) is a construction, as advocated by usage-based models. The problem, however, is to decide what is sufficiently frequent (cf. Gilquin 2015, 53–54). For want of a fixed frequency threshold, some authors have selected the most frequent items as plausible constructions, for example the most frequent part-of-speech (POS) sequences (n-grams) in Cappelle and Grabar (2016) and in Gilquin (2018). Other corpus-based approaches have been used to assign constructional status. Hampe (2011), for example, relies on collostructional analysis, which investigates the strength of association between one or several patterns and some slots within the pattern(s), to postulate the English Denominative Construction ([X SAY [Y BE Z]]).

Different types of experiments have also helped demonstrate the status of constructions. Among these, sentence-sorting experiments have been widely used (e.g. Bencini and Goldberg 2000). They consist in asking participants to sort sentences on the

basis of their overall meaning. If participants predominantly sort the sentences by focusing on their pattern rather than the individual words they consist of, this is a good indication that the pattern is a construction. Syntactic priming, by showing that subjects tend to reproduce the same syntactic structure as they have just been exposed to, can also point to the existence of constructions (e.g. Hare and Goldberg 1999). Some studies have even sought to examine brain activity to make claims about the existence of constructions. Cappelle et al. (2010), for example, rely on magnetoencephalographic evidence to argue that certain phrasal verbs are stored as units in the mind. Interestingly, experimentation has also served to test the reliability of linguistic frequency as a pointer to mental storage (cf. Bybee and Hopper 2001).

These sources of evidence for the constructional status of a pattern may be said to form a cline, from the most indirect (linguistic principles) to the most direct (brain activity) types of evidence about mental representation. While the most direct types of evidence would of course be desirable, not every person interested in linguistic matters may have the possibility of turning to neuroscience or setting up an experimental design to identify constructions. In the next section, we consider how bursts of writing, visible in a special kind of corpus, could give an indication of mental storage and thus help with the identification of constructions.

3. Bursts of writing

3.1. A detour through speech

Before introducing the notion of ‘burst’ in writing and showing how it could be relevant to constructions, it may help to make a detour through spoken language. In Gilquin (forthcoming), it is argued that pauses in speech can help identify constructions, a claim which relies on Dahlmann and Adolphs’s (2009, 126) description of pauses as “indirect indicators of prefabricated language and holistic storage”. If the segment of speech uttered between two pauses is holistically stored, it is a clear sign that this segment might be a construction (provided it corresponds to a pairing of form and meaning). Gilquin (forthcoming) shows, on the basis of results from Gilquin (2018), that the [Prep Det N] pattern is more often interrupted by filled pauses in learner English than in native English, which suggests that the prepositional phrase is more entrenched as a construction in native speech than in learner speech, and that learners may rely more on lower-level constructions like individual words or parts of speech. We will see below how this approach could be transposed to writing.

3.2. Writing fluency and bursts

While the term ‘fluency’ is most typically associated with speech, writing can, just like speech, be more or less fluent. Fluent writing involves writing smoothly and quickly, producing a large number of words in a small amount of time, not pausing too often and

not making too many revisions, among other things (see, e.g., Chenoweth and Hayes 2001; Abdel Latif 2013). The notion of pause thus applies to writing as well, corresponding to a period of time during which the writer is not actually writing.

Chenoweth and Hayes (2001) refer to the segment of text written between pauses as a 'burst'.² Relying on think-aloud protocols, which required students to think aloud while writing a text (one in their L1 and one in an L2), they identify bursts in the verbal reports and compare them for the L1 and L2 texts. They find out that bursts are longer for the L1 texts than for the L2 texts, and also that more proficient L2 writers tend to produce longer bursts. Think-aloud protocols involve spoken data describing the writing process. The use of computers and new technologies, however, have made it possible to approach the writing process more reliably. In particular, keylogging, which records all the keys struck on a keyboard, provides a very accurate representation of the writing process, including all deletions, insertions, substitutions and pauses. Leijten and Van Waes (2013) show how this technology can be used to study bursts. Their tool, Inputlog, automatically counts and measures bursts, once a pause threshold has been set.

3.3. Bursts as constructions

Working on the same assumption as that of Dahlmann and Adolphs (2009) for speech, we may posit that bursts of writing potentially correspond to constructions that are mentally stored, since they are produced in one go, with no interruption between the elements they consist of. Cislaru and Olive (2018), in a thorough analysis of bursts in

keylogging data for French, convincingly argue for the status of bursts as language units, and more precisely units of language production (as opposed to reception; see also Cislaru and Olive 2017). They underline the spontaneous nature of these units, which are the result of “auto-segmentation” by the writer (Cislaru and Olive 2018, 37), rather than being imposed by the analyst. They also demonstrate on the basis of concrete examples from their data that these segments may display formal and semantic regularities, and that certain elements within bursts, the so-called ‘junction attractors’, create strong internal logic, thus contributing to the status of bursts as language units. In fact, Cislaru and Olive (2018, 83) acknowledge that their analysis is inspired by CxG (among other linguistic theories), and they regularly use the term ‘construction’ to describe some of the patterns found in their data, not always necessarily in a CxG sense, but at least in a sense that is compatible with the theories of CxG, e.g. the *N de N* construction (‘N of N’, *ibid.* 113), the *X et Y* construction (‘X and Y’, *ibid.* 159) or the *-ment* suffix (‘-ly’, *ibid.* 236). They even allude to the possible storage of bursts in the mind when they point out that “the spontaneity of burst production may, until proven otherwise, indicate a certain degree of automatisation, which would result in retrieval from long-term memory rather than the implementation of text generation processes” (Cislaru and Olive 2017, 15; my translation).

The above suggests that, provided some form-meaning pairing can be established, a burst of writing could correspond to a mentally stored construction. This, however, should not be seen as a one-to-one correspondence. A burst may include several constructions, if only because constructions can be analysed at different levels, with a construction at a higher level consisting of several constructions at a lower level.

Conversely, a construction may spread over several bursts (see also Cislaru and Olive 2018, 191ff.), for example because the writer was interrupted and could not complete the form-meaning pairing in one go. Pauses in the writing process are therefore indicators of constructions, rather than evidence for them. Corroboration is ideally required from large amounts of process data (to see whether a given pattern most frequently corresponds to a burst or not) and/or from other types of empirical data (see Section 2).

4. Exploration of bursts in PROCEED

PROCEED is a process learner corpus, that is, a corpus representing the writing process of language learners (Gilquin 2021).³ The corpus is made up of argumentative texts written by learners of English and, for each of these texts, a keylog file recorded by means of Inputlog (Leijten and Van Waes 2013) and a screencast video showing the screen activity during the writing process. Only the keylog files are used here. Many of the learners who contributed L2 English data to the corpus additionally wrote an argumentative text in their L1 (mostly French, so far); process data were also collected for these texts. While the corpus includes rich metadata describing each learner's profile (mother tongue, exposure to the target language, but also cognitive measures like working memory or fluid intelligence), these are not exploited for the present purposes.

The PROCEED data used for this study were collected in winter 2018 among students majoring in English at a Belgian university, with an intermediate to advanced proficiency level in English. Only learners whose L1 was French and who had contributed

data both in English and in French have been considered. Bursts have been identified by means of Inputlog, with a minimum threshold of 2000 ms for pauses, following common practice (e.g. Chenoweth and Hayes 2001, Leijten and Van Waes 2013, Cislaru and Olive 2018). Taking the whole sample into account, for a total of 42 texts in English (14,420 words in the finished texts) and 42 texts in French (18,001 words in the finished texts), it appears that bursts are on average less frequent in L1 (2.71 bursts per minute) than in L2 (3.16 bursts per minute) and twice as long in L1 (42 characters) as in L2 (21 characters). This could point to smaller, possibly lower-level constructions in the L2 construction than in the L1 construction, as also suggested in Gilquin (2018).

Examining the actual bursts produced, which has been done for five texts in each language (written by the same students), confirms the presence of many short bursts in the L2 data, sometimes consisting of a single word, e.g. (1), which is made up of three short bursts: *This is, quite and rare.*⁴

(1)

```
{22856}[CAPS LOCK]I[BACK]T[CAPS LOCK]his·is·
{2432}quite·
{2488}rare·
```

In (2), we see how a structure that is usually considered a construction in CxG, the causative construction, is produced in three bursts, which might indicate that the causative construction is not stored in the writer's mind, but only its individual components.

(2)

(...) before·making·this·dream·
{2368}come·
{2232}true· (...)

However, we also notice longer bursts, some of which coincide with structures that one would intuitively describe as constructions, e.g. *in fact; for example; was shocked; first of all; at that moment; the same mistakes; thanks to the law itself; it tells the story of human actions; if you want to change something*. As expected, these can be even longer in L1, up to complete sentences, cf. *il raconte la rébellion des animaux d'une ferme* ('he recounts the rebellion of farm animals'); *Nous savons que ce régime est à la base du régime communiste encore en vigueur en Russie aujourd'hui* ('We know this regime is at the basis of the communist regime still in force in Russia today'); *Bien sûr nous pourrions dire que le voisin A est égal au voisin B mais ces deux voisins sont-ils égaux à l'habitant d'une autre région plus onéreuse ?* ('Of course we could say that neighbour A is equal to neighbour B but are these two neighbours equal to the inhabitant of another more expensive region?'). This tendency to segment the writing process into recognisable constructions is particularly striking in the data of two of the students, both in their L1 and L2 texts. Compare the following extracts produced by one and the same student:

(3)

{10040}[RSHIFT]But·despite·this·fact,·
{12880}and·
{2192}with·the·full·conscience·of·
{10760}what·will·happen,·
{4672}the·[BACK][BACK][BACK][BACK]nearly·the·hole·world·
{19208}would·[BACK][BACK][BACK][BACK][BACK][BACK]will·[BACK][BACK][BACK]ll·go·to·the·
[BACK][BACK][BACK][BACK]battle·again·
{23800}ab[BACK]·bit·more·than·20·years·
{3048}later[RSHIFT].·

(4)

{6424}[RSHIFT]Voici·donc·une·proposition·de·solution·:
{8440}occupons-nous·d'éduquer·au·mieux·les·générations·futures,·en·se·focalisant·sur·k
[BACK]les·zones·les·plus·défavorisées,·
{18288}ne·cherchons·pas·à·tr[BACK]a,n[BACK][BACK]nsformer·des·esprits·formatés·par·la·
violence·
{2616}·:·:·changeons·le·monde·
('{6424}Here then is a proposal for a solution:
{8440}let's concern ourselves with educating the future generations as well as we can, focusing
on the most disadvantaged areas,
{18288}let's not attempt to transform minds shaped by violence
{2616}: let's change the world')

Despite a higher level of segmentation in the L2 example (3), we notice some constructions like a PP (*despite this fact*), an NP (*nearly the whole world*), a relative clause (*what will happen*) and an intransitive motion construction (*will go to the battle again*). The L1 example (4) starts with a presentative construction (*Voici... 'Here is...'*); following this, three imperative forms, together with their arguments, each make up a different burst (*occupons-nous..., ne cherchons pas..., changeons... 'let's concern ourselves..., let's not attempt..., let's change...'*).

At the same time, example (3) shows that some bursts do not correspond to traditionally recognised constructions, as appears from the separation of *a bit more than 20 years* and *later*. Especially interesting is the third burst, ending with *of*, a preposition which one might have expected to go with the next burst. The PROCEED data include

several cases of bursts ending with a function word which, structurally speaking, belongs to the next element, e.g. *a repetition of / killing those...; ...the addition of / a bunch of / factors...; ...represents well the / fact that...; a way to / make...; ...the past doesn't / restrain from repeating; Some event may / seem the same*. This phenomenon can be observed, though less frequently, in the L1 data, cf. *de remplacer le / système* ('to replace the / system'); *...nous permet de / subvenir...* ('...allows us to / support...'); *inventé par les riches pour / essayer de* ('invented by the rich to / try to'). Such cases might be an illustration that not all bursts correspond to constructions (Section 3.3), but they might alternatively be seen as an indication that mentally stored constructions may differ from the kind of constructions that are traditionally recognised (see also Cislaru and Olive 2018, 89ff.). Thus, structurally speaking, prepositions are the head of PPs, and as such we would expect them to form a construction with the rest of the PP. However, considering the above examples, it would also make perfect sense that *of* is stored together with *repetition* or *addition*, for example, as this is part of the valency of these nouns. Similarly, *a bunch of* could be expected to be stored holistically (also for reasons that have to do with frequency and phraseology), despite the fact that *of* is the head of the PP *of factors*. Such structures may remind one of the POS-based n-grams that Cappelle and Grabar (2016) propose to extend in order to form syntactically complete units, for example the addition of an NP after the pattern [X_{noun} *at/in/to/... the* Y_{noun} *of*] or of a nominal after [*at/in/to/ ... the* X_{noun} *of the*]. Although this may be useful from a pedagogical point of view (which is the perspective adopted by the authors), this may not reflect the constructions as they are stored in language users' minds. In fact, Cappelle and Grabar's (2016) results and the present exploration seem to provide

converging evidence that certain mentally stored constructions may be structurally incomplete. This, naturally, should be confirmed by other means, including experimental and ideally neurobiological. In the meantime, we should remain open to the idea that mentally stored constructions may not necessarily correspond to traditional grammatical structures, especially in L2, and we should not forget that our linguistic analyses should mirror the language user's mind rather than the linguist's, as already underlined by Sandra and Rice in 1995.

5. Conclusion

While CxG studies increasingly use empirical data to study constructions, the identification of the constructions to be studied still largely relies on linguistic principles (form-meaning pairings, non-compositional structures, or structures that have been recognised as linguistic units in traditional grammar). The cognitive implication of the constructional status, namely that constructions should be stored in the mind, often tends to be neglected. This is to be regretted, given the “psychologically realistic account of language” that Cognitive CxG seeks to provide (Boas 2013, 242). This article has shown how writing process data arguably offer privileged access to mental representation and could help with the identification of constructions, in the form of bursts of writing, assumed to be stored as units in the mind.

The exploration of L1 and L2 data from PROCEED has revealed bursts that correspond to patterns usually recognised as constructions, thus potentially confirming their mental storage, and has shown that bursts tend to be shorter in L2 than in L1,

which could signal a predominance of lower-level constructions in L2. It also appears from this exploration that some bursts are incomplete, structurally speaking. It has been argued that these could point to mentally stored constructions that are usually not recognised because they do not coincide with grammatical units. As emphasised by Cislaru and Olive (2018), dealing with bursts sometimes involves moving away from traditional grammatical analyses. This, perhaps, is the price to pay to move closer to psychological reality.

References

- Abdel Latif, Muhammad M. Mahmoud. 2013. "What Do We Mean by Writing Fluency and How Can It Be Validly Measured?" *Applied Linguistics* 34 (1): 99-105.
- Bencini, Giulia M.L., and Adele E. Goldberg. 2000. "The Contribution of Argument Structure Constructions to Sentence Meaning." *Journal of Memory and Language* 43 (4): 640-651.
- Boas, Hans C. 2013. "Cognitive Construction Grammar." In *The Oxford Handbook of Construction Grammar*, ed. by Thomas Hoffmann, and Graeme Trousdale, 233-252. Oxford: Oxford University Press.
- Bybee, Joan, and Paul Hopper (eds). 2001. *Frequency and the Emergence of Linguistic Structure*. Amsterdam: John Benjamins.
- Cappelle, Bert, and Natalia Grabar. 2016. "Towards an N-Grammar of English." In *Applied Construction Grammar*, ed. by Sabine De Knop, and Gaëtanelle Gilquin, 271-302. Berlin: De Gruyter.

- Cappelle, Bert, Yury Shtyrov, and Friedemann Pulvermüller. 2010. "Heating Up or Cooling Up the Brain? MEG Evidence That Phrasal Verbs Are Lexical Units." *Brain and Language* 115 (3): 189-201.
- Chenoweth, N. Ann, and John R. Hayes. 2001. "Fluency in Writing: Generating Text in L1 and L2." *Written Communication* 18 (1): 80-98.
- Cislaru, Georgeta, and Thierry Olive. 2017. "Segments répétés, jets textuels et autres routines. Quel niveau de pré-construction ?" *Corpus* 17. <https://doi.org/10.4000/corpus.2846>
- Cislaru, Georgeta, and Thierry Olive. 2018. *Le processus de textualisation. Analyse des unités linguistiques de performance écrite*. Louvain-la-Neuve: De Boeck Supérieur.
- Croft, William. 1998. "Linguistic Evidence and Mental Representations." *Cognitive Linguistics* 9 (2): 151-173.
- Dahlmann, Irina, and Svenja Adolphs. 2009. "Spoken Corpus Analysis: Multimodal Approaches to Language Description." In *Contemporary Corpus Linguistics*, ed. by Paul Baker, 125-139. London: Continuum.
- Gilquin, Gaëtanelle. 2015. "The Use of Phrasal Verbs by French-Speaking EFL Learners. A Constructional and Collostructional Corpus-Based Approach." *Corpus Linguistics and Linguistic Theory* 11 (1): 51-88.
- Gilquin, Gaëtanelle. 2018. "Exploring the Spoken Learner English Constructicon: A Corpus-Driven Approach." In *Speaking in a Second Language*, ed. by Rosa Alonso Alonso, 127-152. Amsterdam: John Benjamins.
- Gilquin, Gaëtanelle. 2021. "Hic Sunt Dracones: Exploring Some *Terra Incognita* in Learner Corpus Research." In *Variation in Time and Space: Observing the World*

- through Corpora*, ed. by Anna Čermáková, and Markéta Malá, 65-86. Berlin: De Gruyter.
- Gilquin, Gaëtanelle. Forthcoming. "Constructing Learner Speech: On the Use of Spoken Data in Applied Construction Grammar." In *Pedagogic Construction Grammar: Data, Methods, and Applications*, ed. by Hans Boas. Berlin: De Gruyter.
- Goldberg, Adele E. 1995. *Constructions. A Construction Grammar Approach to Argument Structure*. Chicago: The University of Chicago Press.
- Goldberg, Adele E. 2003. "Constructions: A New Theoretical Approach to Language." *Trends in Cognitive Sciences* 7 (5): 219-224.
- Goldberg, Adele E. 2006. *Constructions at Work. The Nature of Generalization in Language*. Oxford: Oxford University Press.
- Hampe, Beate. 2011. "Discovering Constructions by Means of Collostruction Analysis: The English Denominative Construction." *Cognitive Linguistics* 22 (2): 211-245.
- Hare, Mary L., and Adele E. Goldberg. 1999. "Structural Priming: Purely Syntactic?" In *Proceedings of the Twenty First Annual Conference of the Cognitive Science Society*, ed. by Martin Hahn, and Scott C. Stoness, 208-211. Mahwah: Lawrence Erlbaum.
- Hilpert, Martin, and Samuel Bourgeois. 2020. "Intersubjectification in Constructional Change: From Confrontation to Solidarity in the *Sarcastic Much?* Construction." *Constructions and Frames* 12 (1): 96-120.
- Leijten, Mariëlle, and Luuk Van Waes. 2013. "Keystroke Logging in Writing Research: Using Inputlog to Analyze and Visualize Writing Processes." *Written Communication* 30 (3): 358-392.

Sandra, Dominiek, and Sally Rice. 1995. "Network Analyses of Prepositional Meaning: Mirroring Whose Mind – the Linguist's or the Language User's?" *Cognitive Linguistics* 6 (1): 89-130.

Author's address

Gaëtanelle Gilquin

Université catholique de Louvain

Centre for English Corpus Linguistics

Place Blaise Pascal 1, bte L3.03.33

B-1348 Louvain-la-Neuve

BELGIUM

gaetanelle.gilquin@uclouvain.be

¹ This section is concerned exclusively with how constructions are identified, i.e. how their status as constructions is determined, not how they are analysed. Thus, Hilpert and Bourgeois (2020) use the linguistic criterion of non-compositionality to justify the constructional status of the *sarcastic much?* pattern, but they then rely on usage data to analyse the construction and describe its features.

² They also take revision (corresponding to a grammatical discontinuity) as a signal for the end of a burst and make a distinction between P-bursts (ending with a pause) and R-bursts (ending with a revision). While bursts demarcated by revisions could also potentially point to constructions, they will not be discussed here.

³ See <https://uclouvain.be/en/research-institutes/ilc/cecl/proceed.html>.

⁴ The figures between curly brackets represent the duration of the pauses in milliseconds; the indications between square brackets show the actions carried out (deletion, capitalisation, etc.).