A Dataset for Studying Idiom Processing with EEG

Philippe Blache, Stéphane Rauzy, Deirdre Bolger, Chotiga Pattamadilok, Sophie Dufour
Aix Marseille University, CNRS, LPL, Aix-en-Provence, France
blache@lpl-aix.fr, stephane.rauzy@lpl-aix.fr

Abstract

We propose in this paper the description of a new dataset aiming at implementing EEG experiments on sentence processing. The resource contains a set of idiomatic sentences together with the corresponding non-idiomatic control sentences. Moreover, in order to study different ERP effects for idiom processing, we also introduce in this original material controlled syntactic violations. As an application, we briefly present an EEG experiment and its results.

Keywords: Idioms, dataset, EEG, syntactic violation

1. Introduction

Studying neural correlates of sentence processing is a difficult task and requires the elaboration of specific material, in which different types of information are controlled (frequency, predictability, syntactic complexity, etc.). Many works focus on phenomena precisely associated with a position or a word, such as the analysis of semantic or syntactic violations introduced by a specific word (Kutas and Federmeier, 2011; Pulvermuller et al., 2008). It is however more complex to study larger phenomena, involving entire constructions (Fillmore, 1988; Goldberg, 2003), with effects at different positions in the sentence. This is still a scientific and methodological lock, and we need to imagine linguistic contexts in which it becomes possible to predict effects at the syntactic level instead of the lexical one. Idiomatic constructions offer such a frame: they can be identified on a word-by-word basis, but are known to be processed globally (Molinaro and Carreiras, 2010; Rommers et al., 2013; Vespignani et al., 2010; Boulenger et al., 2012).

Idioms constitute a prototypical construction (Sag et al., 2002): when recognized, the complete construction (including the meaning as well as possible restrictions on the morphology and the syntax) becomes available. In our work, we intend to analyze *brain activity* in response to a syntactic violation introduced into an idiom. We compare event-related potentials (ERP) in different conditions: idioms vs. control sentences, with or without a syntactic violation. Our goal is to test the hypothesis stipulating that the difficulty of processing the violation is compensated by the activation of the idiom.

This experiment relies on a controlled dataset, made of French sentences in which all information required for implementing such work has been controlled. This constitutes a new resource of 240 sentences, half of them containing idiomatic constructions, the other being corresponding control sentences. Different types of specific information have been encoded such as the familiarity of the idiom, its recognition point as well as information on the type of violation used for this specific study.

2. Linguistic data

A first list of 1,220 French idiomatic expressions have been created from different existing lists available on the web. From this set, a sublist of 170 idioms fulfilling different criteria (familiarity, positions of the constraints violation and its detection) have been extracted by 4 experts. For each idiom, the recognition point *RP* (the word starting from where the idiom is completely recognized) is located. Idioms with "late" *RP*, located at the end of construction, are eliminated, no place being left for introducing violation.

Idiom:	`coûter les yeux de la tête`
Word-by-word:	*to cost an arm and a leg*
Meaning:	*to be very expensive*
English equivalent:	*to cost the eyes in one's head*

Figure 1: Example

In a second stage, this list has been presented to 40 naive participants. Their task was to read the beginning of each idiom (from the first word to the recognition point) and complete them. For each idiom (I), a support sentence (SS) has been built. In order to encapsulate the idiom and to avoid specific effects at the beginning and the end of the sentence, all SS start with a proper noun in a subject position and last with a sentence complement, added after the idiom, in order to let time enough for the EEG signal we want to observe to be realized. The following example illustrates such an idiomatic construction. Starting from the idiom:

(I) `avoir une idée derrière la tête`
(*to have something in mind*)

with recognizing point:

`(RP) derrière`

we build the sentence :

(SS) `Paul a une idée derrière la tête depuis ce matin.`
(*Paul has something in mind since this morning.*)

Idiom selection: From the support sentence (SS), we created the priming stimulus which span from the beginning of sentence to the recognizing point (that is included), as illustrated in the example:

`(PSS) Paul a une idée derrière ...`

The list of the 170 priming stimulus was proposed to a cohort of students (36 for the first 120 items, 25 for the remaining 50 items). It was asked to complete the sentence (without any help). The completion results were analyzed making it possible to calculate a "*familiarity*" measure for each idiom, spanning from 0 (no students can complete correctly the priming stimulus) to 1 (the entire cohort was successful in completing the priming stimulus). The familiar-

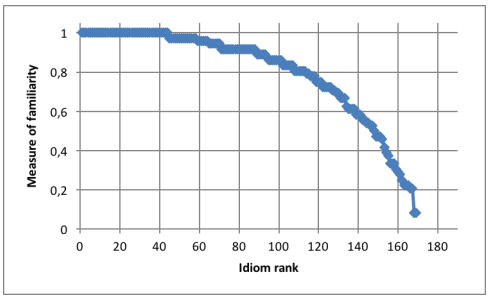

Figure 2: Measure of familiarity

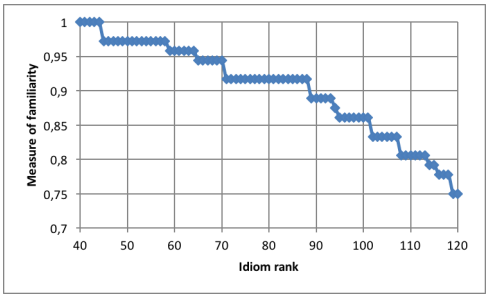

Figure 3: Familiarity of the first 120 idioms

ity measure is illustrated figure 2, the idioms being ranked by decreasing order.

In our list, 44 idioms have been successfully completed by all the participants, the first 100 idioms have a familiarity measure greater than 86% (with means familiarity of 0.96), and the first 120 idioms a familiarity greater than 75% (with a mean familiarity of 0.934) as shown figure 3.

Other idioms receive a lower measure of familiarity for different reasons. One is that the completion of the priming stimulus is ambiguous, as in the following example (which obtains a score of 0.5):

`avoir un verre dans le nez`
(*to be drunk*)

versus the non-idiomatic (but frequent) sentence:

`avoir un verre dans la main`
(*to have a glass in the hand*)

Another reason is that the idiom can be obsolete for the cohort. For example, a measure of familiarity of 0.21 is observed for the idiom:

`boire le calice jusqu'à la lie`
(*To drink from the bitter cup*)

The final list of selected idioms contains the best 120 ranked familiar idioms (with a mean familiarity measure of 0.934).

Control sentences: For each idiomatic support sentence, we created an associated *control sentence* (CS) with the same syntactic structure, the same number of words and as far possible the same lexical material. For example, to the idiomatic support sentence:

(SS) `Paul prend son courage à deux mains pour le faire`
(*Paul takes courage to do it.*)

is associated to the non idiomatic control sentence:

SS	`Paul trouve que ça lui coûte les yeux de la tête maintenant`
SS violation	`Paul trouve que ça lui coûte les yeux sur la tête maintenant`
RP	`yeux`
CS	`Paul trouve que ça lui rappelle les plats de la cantine évidemment`
CS violation	`Paul trouve que ça lui rappelle les plats sur la cantine évidemment`

Figure 4: Idiom, recognition point, control sentence

(CS) `Paul prend son paquet à deux bras pour le porter`
(*Paul takes his bundle with two arms to carry it*)

The table 4 recaps the complete set of data built from one idiom, the corresponding control sentence and the violations.

Violations: The violations are introduced either right after the recognizing point (RP+1) or two words after (RP+2). We introduced two types of violation. The first is a gender and/or number violation agreement between the determiner and the noun in the noun phrase or prepositional phrase following the recognizing point. For example, the support sentence:

(SS) `Paul a un cheveu sur la langue depuis toujours`
(*Paul has a lisp since forever*)

is associated the *violated support sentence* (VSS), in which the agreement between the determiner and the noun is violated:

(VSS) `Paul a un cheveu sur` **le** `langue depuis toujours`

Among the 120 items, there is 47 violations of this kind. The second type of violation introduced is the substitution of the preposition following the recognizing point by another one not allowed in practice, as for example:

(SS) `Paul range au fur et à mesure ses affaires`
(*Paul arranges his stuff as and when*)

(VSS) `Paul range au fur et` **en** `mesure ses affaires`

The list contains 64 items with such violation. The remaining 9 items have slightly different violation rules due to their specific syntactic structure, such as:

(SS) `Paul dit à qui veut l'entendre que c'est vrai`
Paul says to whoever wants to hear it it is true.

(VSS) `Paul dit à qui veut` **s'**`entendre que c'est vrai`

In this case, the violation concerns the accusative pronoun which was substituted by a pronominal pronoun. The same types of violation are also introduced in the control sentences.

3. EEG data

As explained above, an idiom is recognized at the *recognition point* that occurs usually 2 or 3 words after the beginning of the idiom. For example, the recognition point for the idiom "*to put all eggs in one basket*" is the noun "*eggs*". At *RP*, the entire construction is activated, making available predictions about the rest of the input. As illustrated in figure 5, scanning a new word of the input is a simple

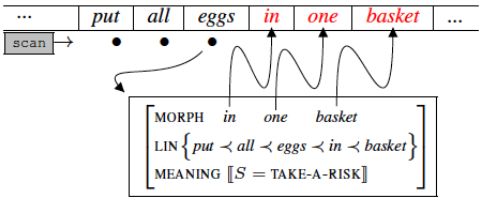

Figure 5: Processing the idiom

mechanism, matching the scanned form with the predicted one. This process remains very shallow, with no precise and in-depth unification mechanism, these words after RP being highly predicted.

One first question to be investigated is to examine whether idiomatic constructions elicit specific brain activities, and more precisely what happens before and after the *RP*. Moreover, in the violation condition, our hypothesis is that there exist compensating effects due to the construction: it is expected that the error in the idiom is identified, but not repaired.

We carried out an electrophysiological (EEG) experiment in which participants were presented with 120 French idioms (ID), 60 with violations (IDV) and 60 without (IDNV), and 120 control sentences (CTR), 60 with violations (CTRV) and 60 without (CTRNV). The stimuli were presented, word-by-word, on-screen during EEG recording. The distribution of idiom familiarity and violation type was controlled. As it is classically the case in EEG, the experiment consists in finding in the data specific electric potentials that can be associated to some stimuli. Several such potentials (called *event related potentials*) are known to be associated with language processing. For example, semantic violations are associated with a negative potential occurring 400ms after the stimulus (called N400), prediction comes with a positive potential 300ms after the stimulus (P300), etc.

		RP	*MM*	*MD1*		
IDNV	*Paul fait la pluie*	*et*	*le*	*beau*	*temps*	...
IDV	*Paul fait la pluie*	*et*	*la*	*beau*	*temps*	...
CRTNV	*Paul fait la peinture*	*et*	*le*	*gros*	*travail*	...
CRTV	*Paul fait la peinture*	*et*	*la*	*gros*	*travail*	...

Table 1: The 4 sentences generated for the idiom "*faire la pluie et le beau temps*" ("*to call the shots*") and the studied positions: the recognition point (*RP*), the modified word (*MM*) where the violation is introduced and the detection word (*MD1*) where the violation is detected for the CRTV condition (here, a gender agreement violation between the determiner and the adjective).

From the 120 sentences in their 4 conditions (idiom, idiom violated, control, control violated), we built two complementary lists of 240 items. Each list contains the whole set of 120 idioms (60 violated and 60 non-violated) and their associated 120 control sentences (60 violated and 60 non-violated). If the couple (violated idiom/non-violated control) belongs to the first list, its corresponding couple (non-violated idiom/violated control) belongs to the second list.

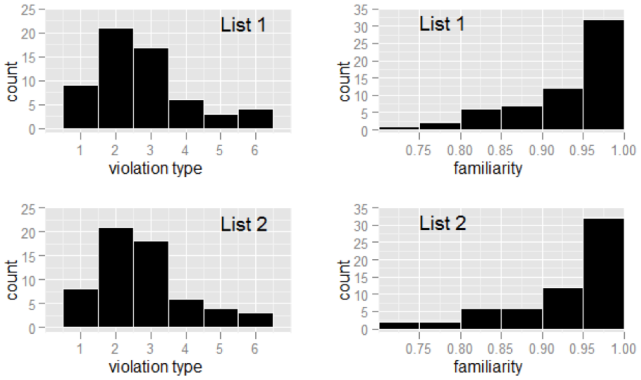

Figure 6: Repartition violation type / familiarity

The two lists have the same distribution of violation (6 different types) for each four conditions (idiom non-violated IDNV, idiom violated IDV, control non-violated CTRNV and control violated CTRV). The two lists have also the same distribution of idiom familiarity (see figure 6).

Figure 1: the distribution of the violation type (left) and of idiom familiarity (right) for the 60 non-violated idioms of the two lists. The mean familiarity is respectively of 0.9345 and 0.9338 for list 1 and list 2.

Subject material input file: The participants were split into two equal subsets, to whom one of the lists 1 or 2 is presented. One single participant can be exposed either to the non-violated idiomatic sentence and its corresponding violated control sentence or to the violated idiom and the non-violated control condition. It never happens that the same participant is asked to read the violated non-violated idiom nor violated and non-violated control sentence. For each participant, the 240 items of the list are randomized and split into six runs of 40 items. For each run, the attention of the participant is checked by inserting 4 sentence questions appealing a yes/no answer to an image presentation. The instruction is to answer *yes* if the sentence has been presented during the last run and *no* otherwise. The answers are balanced in such a way that a given subject has 12 positive and 12 negative answers to give over the experiment.

4. An EEG/ERP study on syntactic violations in idiom comprehension

Subject-level, trial-averaged EEG data was extracted for the three word positions: the *Recognition Point* (RP), the *Modified Word* (MM) where the violation is introduced and the *Detection Word* (MD1) where the violation is detected (for the CRTV condition). A two-tailed cluster-based permutation was carried out on the data for both CTRLs and IDs to compare non-violation conditions (CTRNV and IDNV) and violation conditions (CTRV and IDV).

Recognition Point (RP): As no effect of violation was expected, the violation conditions were collapsed for both CTRL and ID ((CTRNV+CTRV) vs. (IDNV+IDV)).

Statistical analyses revealed a significant ($p \leq 0.025$, two-tailed) N400 difference over centro-parietal electrodes from $\sim$390 to 550ms; CTRL presented a higher N400 ampli-

tude than ID (figure 1). N400 amplitude is generally thought to increase as a function of the difficulty of word retrieval and integration (Kutas and Federmeier, 2011). This observation is in line with previous findings of a reduced N400 in the context of idioms compared to literal sentences (Rommers et al., 2013) and is indicative of higher word-predictability at the RP for idioms compared to controls. A greater P300 effect, posited as an index of prediction processes in idioms (Molinaro and Carreiras, 2010), was observed for ID compared to CTRL. However, this did not reach statistical significance according to the cluster-based permutation test (figure 7).

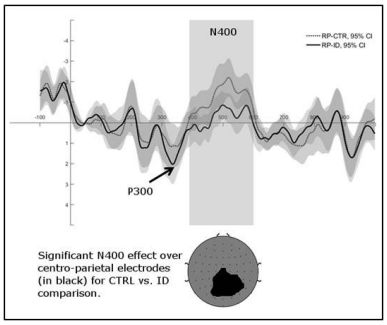

Figure 7: RP Position

Modified Word (MM): Violation effect in CTRL and ID were analyzed separately. As expected, no significant difference was revealed for CTRLs. However, for ID, IDV presented a significantly higher ($p \leq 0.025$) N400 than IDNV (figure 8).

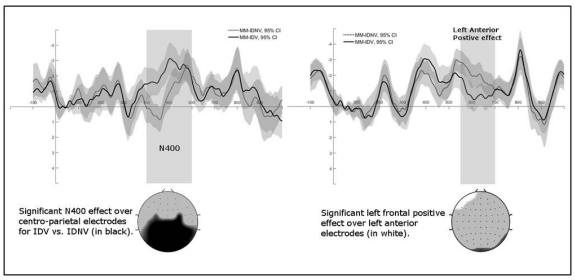

Figure 8: Position MM

This N400 effect indexes the violation of the prediction made at the RP and, so far as it indicates that the violation has already been processed, this effect also implies that, for ID, the reader is already predicting the error that will occur at MD1.

A significant difference ($p \leq 0.025$) between IDV and IDNV in the 550 to 700ms time window over left frontal electrodes (figure 2) was also revealed; IDV presented more positive-going activity compared to IDNV. This observation could be interpreted in light of (Hagoort et al., 1999) suggestion that more frontally distributed P600-like effects may reflect the over-writing of an "*active structural representation*".

Detection Word (MD1): At this position the reader detects the violation introduced at position MM for CTRL. A CTRV vs. CTRNV comparison revealed a significant N400 effect ($p \leq 0.05$) (figure 3, left); this reflects the processing of the control violation.

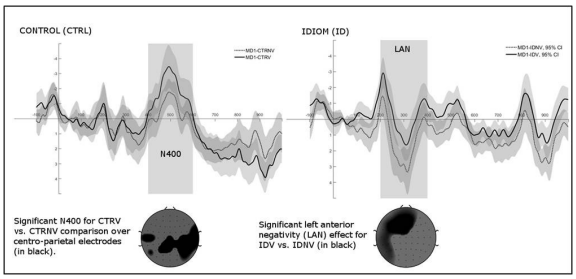

Fig. 3: MD1 position

The N400 effect is followed by increased positive-going activity for CTRV from around 600ms; this P600 effect indicates the processing of the syntactic violation. The ID condition presents a different pattern of results. The N400 was very much reduced for both IDNV and IDV and no significant N400 difference was observed as a function of the violation. However, an IDNV vs. IDV comparison revealed a significant ($p \leq 0.025$) difference over left frontal electrodes from around 200ms to 400ms, (figure 3, right) with IDV presenting more negative-going activity than IDNV. The temporal and spatial focus of this effect suggests a LAN (Left Anterior Negativity) which has been posited as reflecting syntactic processing (Friederici et al., 1996; Kluender and Kutas, 1993) rather than semantic integration.
These results validate the different hypothesis mentionned above. By showing a higher positivity (reduced N400, higher P300) starting from the recognition point (*RP*), the ERPs validate the facilitator effect after *RP* due to the prediction of the entire construction. At the modified word position (*MM*), as predicted by the model, the violation in the idiomatic construction is detected (small N400). Moreover, the unexpected element is recovered (P600), corresponding to our constraint relaxation hypothesis. The analysis of the detection word position (*MD1*) reveals clearly a specificity of violation in idiomatic constructions. In the IDV condition, the violation is already predicted starting from the modified word (*MM*). This explains the fact that no N400 occurs at this point in IDV. In the control condition, the violation is only detected at this point, which explains a high N400, followed by a repair. Finally, as predicted by the model, the violation is not repaired in IDV: the LAN at this position reveals an automatic detection of the violation, but is not followed by a repair that should have generated a P600.

5. Conclusion

The dataset presented in this paper constitutes a complete resource for the study of idiom processing. The EEG experiment done with this resource shows the compensation effect played by the idiomatic construction when faced with a syntactic violation. Such dataset opens new experimental possibilities. On top of providing a controlled material for testing hypothesis on idiom processing, it also opens directions towards new experiments in neurolinguistics for the analysis of syntactic phenomena at the sentence level. In particular, the fact that entire constructions such as idiom can be manipulated makes it possible to implement different experiments involving larger contexts than isolated words or adjacent chunks. The EEG experiment we presented is an illustration of such type of works.

6. Acknowledgements

This work has been supported by grants ANR-16-CONV-0002 (ILCB), ANR-11-LABX-0036 (BLRI) and ANR-11-IDEX-0001-02 (A*MIDEX).

7. Bibliographical References

Boulenger, V., Shtyrov, Y., and Pulvermüller, F. (2012). When do you grasp the idea? meg evidence for instantaneous idiom understanding. *NeuroImage*, 59(4):1–12.

Fillmore, C. J. (1988). The mechanisms of "construction grammar". In *Proceedings of the Fourteenth Annual Meeting of the Berkeley Linguistics Society*, pages 35–55.

Friederici, A., Hahne, A., and Mecklinger, A. (1996). Temporal structure of syntactic parsing: Early and late event-related brain potential effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22:1219–1248.

Goldberg, A. E. (2003). Constructions: a new theoretical approach to language. *Trends in Cognitive Sciences*, 7(5):219–224.

Hagoort, P., Brown, C. M., and Osterhout, L. (1999). The neurocognition of syntactic processing. In C. M. Brown et al., editors, *The neurocognition of language*. Oxford University Press.

Kluender, R. and Kutas, M. (1993). Subjacency as a processing phenomenon. *Language and Cognitive Processes*, 8(4):573–633.

Kutas, M. and Federmeier, K. (2011). Thirty years and counting: Finding meaning in the n400 component of the event-related brain potential (erp). *The Annual Review of Psychology*, 62(1):621–647.

Molinaro, N. and Carreiras, M. (2010). Electrophysiological evidence of interaction between contextual expectation and semantic integration during the processing of collocations. *Biological Psychology*, 83(3):176–190.

Pulvermuller, F., Shtyrov, Y., Hasting, A. S., and Carlyon, R. P. (2008). Syntax as a reflex: Neurophysiological evidence for early automaticity of grammatical processing. *Brain and Language*, 104(3):244–253.

Rommers, J., Dijkstra, T., and Bastiaansen, M. (2013). Context-dependent Semantic Processing in the Human Brain: Evidence from Idiom Comprehension. *Journal of Cognitive Neuroscience*, 25(5):762–776, May.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for nlp. In *Proceedings of International Conference on Intelligent Text Processing and Computational Linguistics*, pages 1–15.

Vespignani, F., Canal, P., Molinaro, N., Fonda, S., and Cacciari, C. (2010). Predictive mechanisms in idiom comprehension. *Journal of Cognitive Neuroscience*, 22(8):1682–1700.