BLRI
Brain & Language Research Institute

Project

Interaction entre complexification et facilitation dans le traitement du langage

Etude de la violation de contraintes dans les idiomes

Philippe Blache[(1)], Sophie Dufour[(1)], Chotiga Pattamadilok[(1)], Carlos Ramisch, Stéphane Rauzy[(1)]

(1) LPL, (2) LIF

Abstract

Présentation

L'étude des facteurs de complexité du traitement du langage consiste à identifier les paramètres, les phénomènes qui induisent une difficulté accrue. De nombreux travaux ont ainsi proposé des modèles de difficulté, en particulier au niveau de la phrase (Gibson, 2000; Warren & Gibson, 2002; Lewis & Vasishth, 2005). Ces travaux mettent à jour des informations variées, relevant de caractéristiques lexicales et morphosyntaxiques (fréquence, longueur des mots, structure morphologique, etc.), syntaxiques (niveaux de profondeur de la structure, éloignement des dépendances), sémantiques ou encore discursives (identification des référents et leur accessibilité). Parallèlement aux effets de difficulté liés à des constructions particulières, d'autres paramètres peuvent également jouer un rôle important dans l'évaluation de la complexité. Il s'agit en particulier des phénomènes de violation : une phrase ou un énoncé violant une règle ou une propriété linguistique induira une complexité de traitement accrue (Sorace & Keller, 2005). Plusieurs approches proposent une estimation quantifiée de la difficulté en exploitant notamment les informations fournies par des analyseurs automatiques. C'est le cas notamment de l'indice de surprise mesurant la prédictibilité d'une forme (Hale, 2001) ou encore de la complexité de la structure d'analyse estimée par la profondeur des piles de stockage d'un analyseur (Schuler et al., 2008).

Par ailleurs, plusieurs études ont montré que ces facteurs de difficulté étaient sensibles à un effet de cumulativité (Keller, 2006) : le nombre de propriétés violées est corrélé avec le niveau de complexité. Le phénomène de cumulativité négative (dans le sens ou les facteurs de difficulté s'additionnent pour rendre le processus de traitement plus complexe) est d'ailleurs dans une certaine mesure pris en compte par les estimations probabilistes issues des analyseurs. Dans le même temps, mais de façon plus limitée, des études ont montré de surcroît l'effet facilitateur que certains paramètres pouvait conférer au traitement. Typiquement, les phénomènes de niveau d'activation permettent de moduler certains facteurs de complexité : le traitement d'une construction peut être facilité en fonction de son contexte (Vasishth, 2003).

La question qui se pose alors est celle de l'estimation de l'interaction existant entre effets de complexification et de facilitation. Des travaux ont ainsi montré que le niveau de grammaticalité d'une phrase pouvait être décrit et évalué en prenant en compte à la fois les propriétés violées et satisfaites. Ce niveau de grammaticalité ainsi estimé est corrélé au jugement d'acceptabilité (Blache et al., 2006). Il y aurait donc, parallèlement à une cumulativité négative, un effet de cumulativité positive qu'il serait possible d'estimer en fonction des caractéristiques linguistiques exprimées en termes de violation et satisfaction de contraintes.

Le but de cette étude est de rechercher une validation expérimentale de ce modèle de complexité. Nous proposons pour cela d'étudier des phrases comportant une construction idiomatique. Ce type de phénomène est en effet connu pour introduire un effet de facilitation puissant (Vespignani et al., 2010; Rommers et al., 2013). Cet effet provient, selon notre modèle, à la fois du nombre et du poids des propriétés satisfaites et non satisfaites. Notre hypothèse est que l'introduction d'une violation syntaxique au sein d'un idiome se trouve compensée par le niveau élevé de facilitation produit par l'idiome. Pour étudier ce phénomène, nous proposons de comparer l'effet d'une même violation syntaxique sur la phrase comportant un idiome par rapport à une phrase contrôle de même type ne comportant pas d'idiome. Selon notre hypothèse, la violation syntaxique sera peu ou faiblement détectée, ce qui devrait se traduire par des effets moindre de détection de violation en termes de potentiels évoqués.

Méthode

Nous proposons de conduire une expérience permettant d'examiner les potentiels évoqués de sujets confrontés à la lecture de phrase contenant des idiomes avec violation syntaxique, par comparaison avec des phrases contrôles.

120 idiomes de type « avoir une idée derrière la tête » ont été sélectionnés. Pour chaque idiome, 4 phrases ont été construites donnant lieu à 4 conditions expérimentales :

1) Une phrase idiomatique sans violation telle que « Paul a une idée derrière la tête depuis ce matin »

2) Une phrase idiomatique avec violation syntaxique introduite après le point de reconnaissance de l'idiome. « Paul a une idée derrière le tête depuis ce matin »

3) Une phrase non idiomatique sans violation ayant la même structure syntaxique que l'idiome : « Paul a une douleur derrière la nuque depuis ce matin »

4) Une phrase non idiomatique avec violation syntaxique introduite au même endroit que dans l'idiome « Paul a une douleur derrière le nuque depuis ce matin »

A noter que les violations syntaxiques portent toujours sur un mot de fonction.

Pour s'assurer que les sujets prêtent attention aux phrases, des essais fillers contenant une phrase suivie d'une image ont été ajoutés (10%). A l'apparition des images, les participants doivent décider si l'image correspond à la phrase qui la précède.

Les phrases seront présentées visuellement au rythme d'un mot toutes les 600 ms (ms (le mot est affiché pendant 300 ms suivi par l'écran noir pendant 300 ms). Les réponses électrophysiologiques seront mesurées au moment de la violation et sur le mot contenu qui suit la violation.

Pour chaque point de mesure, 3 composantes seront analysées

1) ELAN (100-300ms) : reflète la détection automatique d'une violation syntaxique

2) LAN et N400 (300-500 ms) : reflète l'intégration morpho-syntaxique ou lexico sémantique

3) P600 (500-1000 ms) : reflète un processus de réparation ou ré-analyse de la phrase suite à la détection de la violation

Participants

30 participants d'origine francophone, droitiers sans déficits visuels, auditifs, langagiers et neurophysiologiques, âges compris entre 18 et 35 ans seront recrutés.

Situation par rapport aux objectifs du BLRI

Ce projet répond aux objectifs du BLRI de plusieurs façons :

- Il s'agit d'un projet interdisciplinaire, réunissant des linguistes, des informaticiens et des psycholinguistes de deux laboratoires du BLRI (le LPL et le LIF). L'objectif de ce projet est de valider une hypothèse formulée dans le cadre d'une théorie linguistique selon laquelle la complexité du traitement syntaxique dépend de la quantité et de la qualité des informations disponibles (les contraintes satisfaites ou violées ainsi que leur importance). On utilise pour cela une modélisation computationnelle permettant de quantifier cette information. Une estimation plus fine, adaptée au type de construction particulier utilisé dans cette étude (les idiomes), utilisera les techniques développées par les partenaires du projet.

- Ce projet offre la possibilité d'avancer à partir d'un point de vue théorique et formel en linguistique vers une expérimentation permettant de poser la question de la modélisation du traitement syntaxique en adoptant une perspective jusqu'ici peu explorée du point de vue expérimental. Il répond à l'un des objectifs du BLRI de mettre en place de nouvelles expérimentations, partant de questionnements n'ayant pas encore conduit à des études en direction du fonctionnement cérébral. En d'autres termes, il s'agit d'une problématique provenant de questions posées par des linguistes et des informaticiens, sans que ces questions n'aient fait jusqu' à présent l'objet de recherches en direction des bases cérébrales.

- Ce projet permet d'envisager, à partir du même jeu de données, des investigations variées, par exemple en mouvement oculaire ou en MEG.

Demande de soutien auprès du BLRI

1. Aide à la préparation et la passation de l'expérience EEG

    a. Convocation des sujets

b. Passation des expériences

2. Traitement des données

3. Indemnisation des sujets : 30*20€ = 600€

Publications

-

Fiche-résumé contribution CREx

Idiome

The interaction between complexification and facilitation in the processing of language

Investigateurs : Philippe Blache, Sophie Dufour, Chotiga Pattamadilok, Carlos Ramisch, Stéphane Rauzy.

Durée : 24 months

Contribution :

- Running of the EEG experiments
- Preprocessing of the raw EEG data.
- Analysis of EEG – ERPs and statistical analysis
- Preparation of poster presentation and methods and analysis sections pertaining to EEG preprocessing and statistical analysis of ERP data.

Objective : *The aim of this study is provide an experimental validation of a model of complexity. This model is based on the idea of the existence of an interaction between complexification effects and facilitation effects, in which the level of grammaticality of a multiword structure can be evaluated by taking into account both those constraints that are violated and those that are satisfied. This, thus, leads to both cumulative positive and negative effects.*

Passation

The acquisition of EEG data was carried out at the CEP of the LPL. Twenty-five participants carried out the experiment which involved reading sentences presented on-screen word by word. The phrases included both normal sentences and idioms and, for both sentence types, a violation was introduced. This violation was always introduced after the “Recognition Point” of the idiomatic sentence. To

ensure that the participants' level of attention was kept high, questions followed sentence presentation in a random manner (10% of trials); the participants were presented with an image and had to indicate by button-press, if "yes" or "no" the image corresponded to the preceding sentence.

EEG signals were recording using the Active2 Biosemi system with 64 electrodes. The online reference was the left mastoid and the online sampling rate was 2048Hz. Experiment presentation was control via the Eprime software.

Preprocessing

Preprocessing was carried out by the CREx using a preprocessing script prepared by the engineer using the Matlab toolbox, EEGLAB. Figure 1 presents the pre-processing pipeline applied for each subjects. The data of three subjects was excluded from the analysis due to a high level of noise.

The continuous data was segmented on the basis of three conditions corresponding to three word positions within the sentence:

- Recognition Point (RP) – point at which reader recognizes the idiom
- Modified Word (MM)- point at which the violation is introduced
- Detection Word (MD)-point at which the violation is detected by the reader.

These sentence points were isolated for both Control Sentences (CTR) with violation (CTRV) and without violation (CTRNV) and for Idioms (ID), with violation (IDV) and without violation (IDNV). This yielded the following 12 conditions, on which we would focus our analyses:

- CTRNV – RP
- CTRV –RP
- CTRNV – MM
- CTRV –MM
- CTRNV –MD
- CRV – MD
- IDNV – RP
- IDV-RP
- IDNV –MM
- IDV – MM
- IDNV – MD
- IDV - MD

Total epoch length was 1100ms; a baseline of 100ms and a post-stimulus interval of 1000ms. Baseline normalization was applied after segmentation of the continuous data.

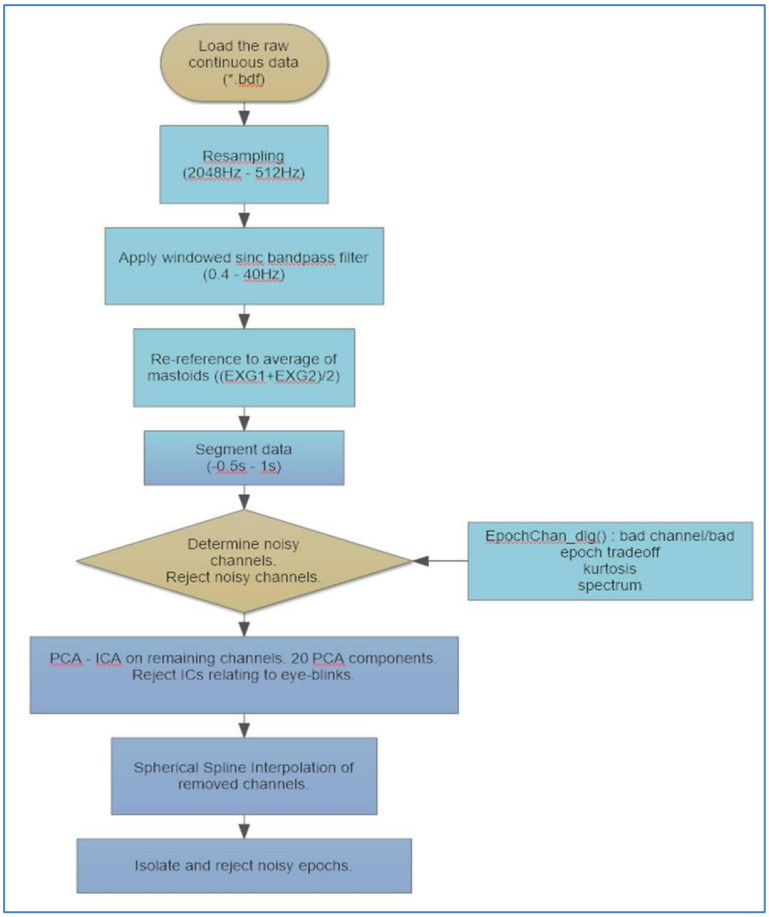

Figure 1: *Pre-processing pipeline applied for each subject.*

Analysis

The grand average ERP data was analyzed with a particular focus on the following comparisons :

- CTRNV-RP vs CTRV – RP (No difference expected as before MM)
- IDNV-RP vs. IDV-RP (no difference expected as before MM)
- CTR (NV+V) vs. ID (NV+V) (a possible difference due to recognition of idiom…memory)
- CTRNV-MM vs. CTRV-MM (no difference expected due to violation, as before MD)
- IDNV-MM vs. IDV-MM (difference expected due to violation of expectancy– N400).
- CTRNV-MD vs. CTRV-MD (difference expected due to processing of violation – N400)
- IDNV-MD vs. IDV-MD (a reduced effect of the violation, compared to CTRLs…maybe a re-analysis).

To analyse the data in a data-driven manner, without making a-priori assumption regarding time-windows of interest or brain-regions of interest, **non-parametric analyses** were carried out on the grand-average ERP data for all sample points of the entire epoch. A **permutation test with fdr correction** was carried for each of the above 7 comparisons for each of the 64 electrodes, 2000 permutations were applied. For each electrode, temporal intervals presenting statistically significant differences ($p \leq 0.05$) whose duration exceeded 10ms were considered. Figure 2 presents the result of this permutation test for CTR (NV+V) vs. IDV (NV+V) for RP. Those intervals presenting statistically significant differences are marked with red (eg. Pz, CP2, CP4…). The greater N400 effect for CTRs

compared to IDs at the RP may be due to a higher cloze-probability for idioms (once recognized) compared to controls (less expectancy) (Roehm, 2007).

***Figure 2**: Result of permutation test with fdr correction for contrast CTR (NV+V) vs. ID (NV+V) for all time samples and all 64 electrodes. Those temporal intervals (>10ms) presenting statistically significant ($p \leq 0.05$) effect of sentence-type are marked with red. The Pz electrode is presented in detail against the topographies (ID-CTR) in steps of 100ms.*

To isolate, not only temporal intervals presenting statistically significant condition-related effects, but also spatial regions over which such effects are concentrated, a cluster-based permutation test was carried out over all time samples of the epoch and over all 64 electrodes. To isolate both positive and negative-going effects, a two-tailed test was applied ($p \leq 0.025$). The cluster-based permutation facilitates the resolution of the multiple comparisons problem by generating clusters based, in this case, on both temporal and spatial proximity and applying the correction within these clusters. This analysis was carried out using the Matlab toolbox, FieldTrip. Figure 4 presents the results of the two-tailed cluster-based permutation test for the IDV-MM vs. IDNV-MM comparison. The topographies of IDV-IDNV are presented over time in 100ms steps and those electrodes presenting statistically significant differences ($p \leq 0.025$) are marked with a black x if effect is negative and a white x if effect is positive.

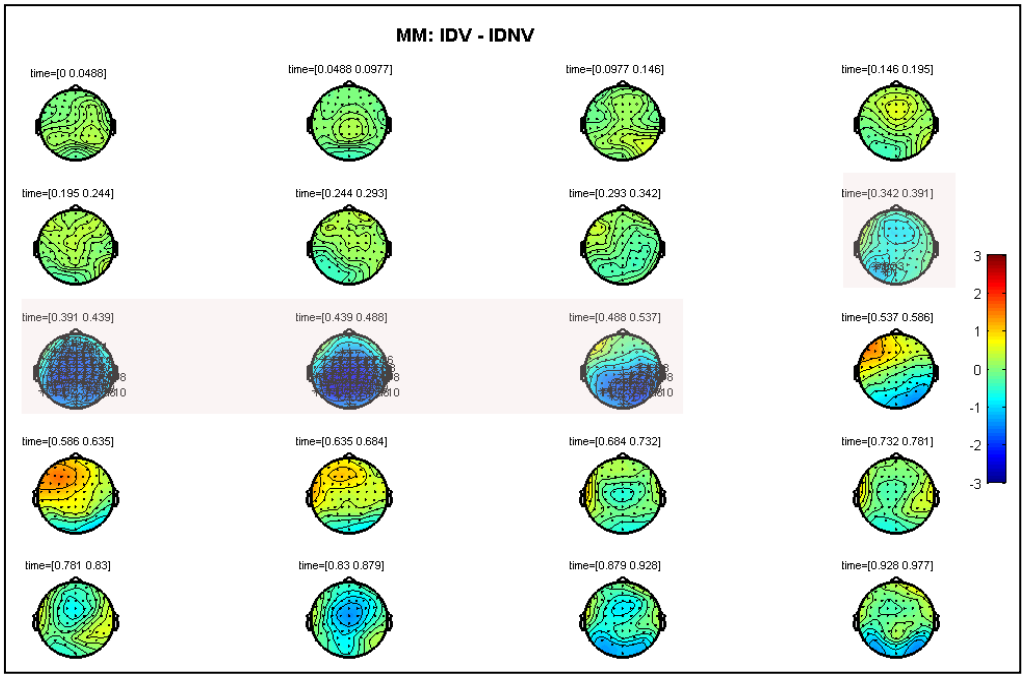

Figure 4: *Result of two-tailed cluster-based permutation test ($p \leq 0.025$) on the IDV vs. IDNV comparison for position MM. The results are presented against the topographies of IDV-IDNV in steps of 100ms and reveal a greater negative effect for IDV compared to IDNV from around 391ms to 537ms over central and parietal regions in particular. This result strongly indicates a greater N400 effect for IDV compared to IDNV, due to the introduction of the violation.*

Diffusion

Poster presentation at Cargèse 2015 :
Presentation_IDIOME_Cargèse44x44.pdf

Article submitted for Conference of ACL (Association of Computation Linguistics) 2017:
Blache_acl2017-1.pdf