A systematic approach to identify lexical makers in French conversations: A pilot study

Klim Peshkov, Laurent Prévot
Aix Marseille Université, Laboratoire Parole et Langage, UMR 7309, Aix-en-Provence, France
klim.peshkov@univ-amu.fr, laurent.prevot@univ-amu.fr

Abstract content

This work has for objective to identify the main lexical discourse makers used in French dialogues. The current stage of the work focuses on French conversations of the CID corpus (Bertrand et al., 2008). Our work relies on a manual segmentation of fine-grained discourse units (DU) from which we extract the most frequent initial and final tokens. We then rely on the strong hypothesis that these positions are the preferred ones for IP-adjuncts sentence elements that are traditionally described as holding a discourse function.

More precisely, we split our discourse units (that are defined thanks to a combination of semantic and pragmatic criteria) into short discourse units (SDU), composed of 3 or less tokens (words), and other (long) discourse units (LDU). Our dataset is made of 12500 'long' and 5572 'short' discourse units.

Our idea is to characterize these words based on joint consideration of their distribution on final and initial positions of both short and long discourse units. We take the lexical items frequently occurring within short discourse units as representative of interactional discourse. In most cases, they correspond to feedback *(mh, ouais, <laughter>, voilà, d'accord, oui, non)* (Bunt, 1994), turn-management signals and filled pauses of disfluencies *(euh)*. Some of them appear in this list because they are components of more complex markers such as *c'est ça*.

We consider jointly SDU and LDU initial and final tokens by looking for each token at the ratios $Init\_Ratio = \frac{\#init-LDU}{\#init-SDU}$ and $Fin\_Ratio = \frac{\#final-LDU}{\#final-SDU}$. Different categories clearly emerge as can be seen in figure 1.

High $Init\_Ratio$ (> 8) corresponds to several categories: (i) syntactic device for introducing subordinate or independent clauses (and thus excellent DRD candidates), we list them here together with tentative default discourse relation associated (in SDRT (Asher and Lascarides, 2003) framework) (*où, quand*: BACKGROUND/CONTINGENCY, *parceque, puisque*: EXPLANATION/CAUSAL, *c'est-à-dire, même*: ELABORATION), *sinon*: CONTRAST, *par exemple*: EXEMPLIFICATION), *si, alors, sinon*: CONDITION/LOGICAL), *puis*: NARRATION). Those are mixed with markers of specific spoken constructions such as clefts *(moi)*, relative pronouns introducing relative clauses *(qu', que)*[1], pronouns and determiners that frequently occupy subject positions and finally interesting discourse marking usage of other words: (*genre*: ELABORATION).

Intermediate $\frac{\#fin\,in\,long}{\#fin\,in\,short}$ ratio ($1 < ratio < 8$) typically delineates words holding both connective and more interactional functions. We list tentative discourse relations but are fully aware of the special difficulty of disambiguating those between their discourse marking and interactional usage (*et*, *mais*: CONTRAST, *donc, alors*:RESULT/CAUSAL, *après*:NARRATION, *enfin*:REFORMULATION).

Concerning high Fin_Ratio (> 8), aside from frequent nouns which do not occur in short discourse units, we identify specific spoken final particles (*quoi, hein, en fait*) or part of it (*vois / sais* of *tu vois / sais*), together with parts of specific constructions (*non plus*).

As a first step of analysis, low Fin_Ratio and $Init\_Ratio$ ($ratio < 1$) can be considered simultaneously. Overall they consist of feedback related words: *mh, d'accord, voilà, ouais, oui, super* and evaluative sentence adverbs *effectiment, simplement, évidemment*.

We cannot do due justice here to the impressive amount of existing work on these markers. We hope however to discuss at the workshop our simple methodology that can be applied to various corpora and other languages and that can be used to prioritize the work on spoken DRDs.

[1] It is often possible to identify discourse relations for these relatives.

1. References

Asher, N. and Lascarides, A. (2003). *Logics of conversation*. Cambridge University Press.

Bertrand, R., Blache, P., Espesser, R., Ferré, G., Meunier, C., Priego-Valverde, B., and Rauzy, S. (2008). Le cid-corpus of interactional data-annotation et exploitation multimodale de parole conversationnelle. *Traitement automatique des langues*, 49(3):1–30.

Bunt, H. (1994). Context and dialogue control. *Think Quarterly*, 3(1):19–31.

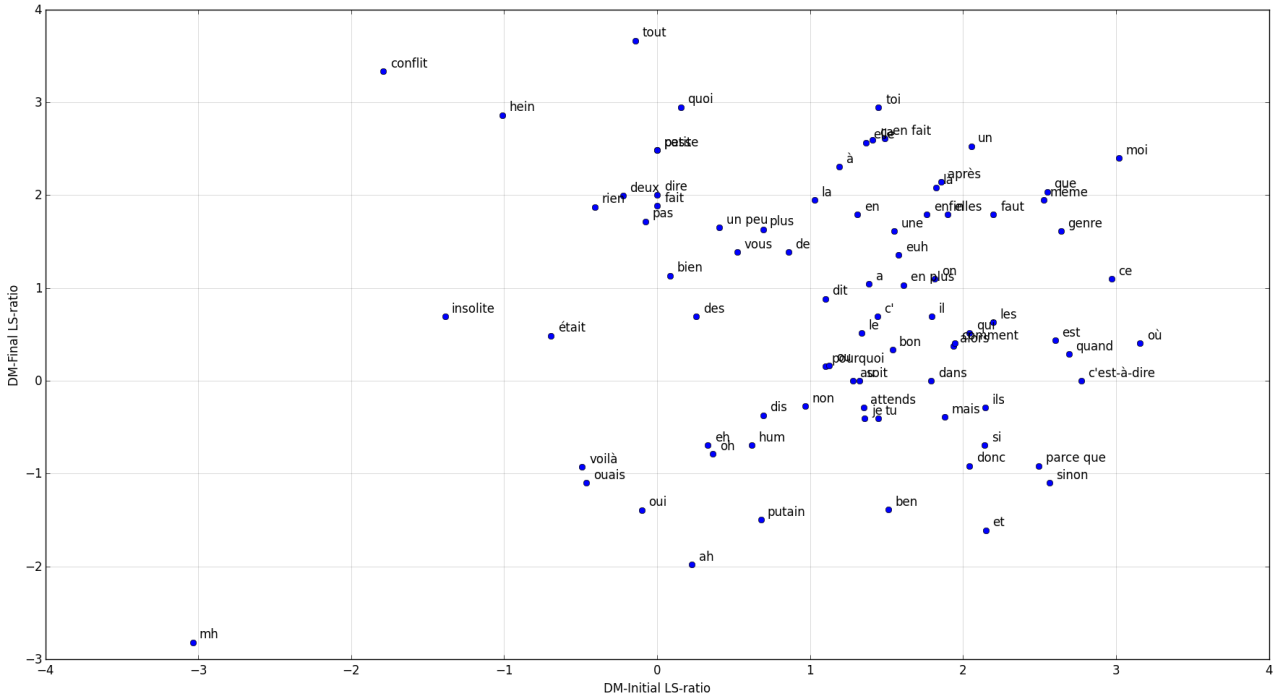

Figure 1: Log of Long/Short-ratios of frequent tokens in initial/final positions