Grouping conversational markers across languages by exploiting large comparable corpora and unsupervised segmentation

Laurent Prévot[1,2], Matthieu Stali[1], Shu-Chuan Tseng[3]
[1] Aix-Marseille Univ, LPL, Aix-en-Provence, France
[2] Institut Universitaire de France, Paris, France
[3] Institute of Linguistics, Academia Sinica, Taipei, Taiwan

Abstract

This work approaches *Conversational and Discourse Markers* (hereafter DM) from a radical data-driven perspective grounded in large comparable corpora of French, English and Taiwan Mandarin conversations. The key features of our approach are (i) to account for lexicalization as a by-product of unsupervised segmentation applied to our corpora, (ii) to exploit simple metrics for clustering DM (both within a language and within multilingual clusters). We explore the benefits and the drawbacks of such a radical approach to DM. In particular we compare the DM clusters obtained from traditional segmentation into tokens (as given by manual transcription of the corpora) vs. unsupervised segmentation. The metrics on which we ground the clustering experiments are based on contrast between (i) short vs. longer utterances distribution and (ii) position within longer utterances.

1. Introduction

Leaving aside some interesting descriptive studies, there are not many attempts to perform systematic and quantitative comparative analysis of social interactions (such as conversations and task-oriented dialogues) from a linguistic perspective. Language resources and natural language processing tools still rely on written canonical data. In the context of studying, comparing and exploiting social interactions; in which speech is fiercely spontaneous and exhibits its own patterns; appears to be a major handicap. Once situated within a multilingual or translational task, it becomes even more difficult to handle by adding the bias towards written canonical data of each language before being able to consider the multilingual or translational aspects themselves. Thus, we propose here to adopt a relatively shallow and data-intensive approach to consider directly the spoken data without passing through resources and tools created for canonical written data.

Comparable corpora are extremely useful for a range of Human Language Technology tasks but also for exploring phenomena across languages. In this paper we are developing a data-driven approach to study discourse and interactional markers (hereafter DM) in a comparative way thanks to large conversational comparable corpora. Our work aims at identifying and grouping discourse markers into homogeneous classes through a purely bottom-up approach carried out on large corpora. Studying discourse markers has a long history in linguistics and corpus linguistics (see Section 2.) but our approach combine some methodological choices that makes it original. This approach relies on rather large comparable conversational corpora across the languages scrutinized (introduced in section 3.). Moreover those corpora have to be transcribed. More precisely the two key ingredients are (i) to explore unsupervised segmentation of our data sets as explained in 4.1. ; (ii) to explore a set of distributional measures of the word-like units for characterizing them (See 4.2.). Finally, in our experiments, standard clustering techniques are used to obtain groups of clusters that we try to label with categories in section 5..

2. Discourse Markers

Discourse markers, such as *like* and *well* in English to quote a few, are key elements in conversations which help speakers build their speech's structure. The main issue when studying DMs lies in the lack of consensus and thus in the various definitions and denominations that can be found among works in the literature related to conversational speech. We can mention the following terms, being the most frequently used: *discourse markers* (Schiffrin, 1988; Fraser, 1999); *pragmatic markers* (Furko, 2009; Garric and Calas, 2007), *discourse particles* (Schourup, 1985; Fischer, 2006), *spoken particles*(Fernandez, 1994; Fernandez-Vest, 2015) and *discourse connectives* (Roze et al., 2010; Lenk, 1998).

Even though we can understand why a categorization task for DMs remain difficult given their poly-functionality and the various stages of functional multi-word expressions' lexicalization, scholars would usually agree on several main aspects. DMs' primary functions are described as being related to a relatively defined set of functions: turn-taking system, discourse relations cuing, discourse structuring, interpersonal relationships marking, speech management or politeness (Fischer, 2006).

Recently, linguists have been interested in automatically identifying DMs for translation purposes. Some results have shown there were discrepancies between bilingual dictionaries translations and the semi-manual annotation ones for a given pair of DMs from two different languages (Roze and Danlos, 2011). Other works include *The TextLink project*[1] which is specifi-

[1] `http://textlinkcost.wixsite.com/textlink`

cally analyzing this aspect, by focusing on discourse-annotated corpora to allow cross-linguistic studies of discourse. The corpus based method seems an adequate tool for caterorizing DMs as it unites a theoretical task consisting in setting parameters of definition variables with an empirical study on spontaneous speech corpora (Crible et al., 2015).

3. Data

The comparable corpora we used for this experiment were : the CoFee collection of corpora (Prévot et al., 2016) (made of CID (Blache et al., 2009), MapTask(Gorisch et al., 2014) and DVD(Prévot et al., 2016)) together with DECODA corpus for French ; Switchoard transcripts for English (Godfrey et al., 1992) ; and Academia Sinica conversational corpora (MCDC, MTCC, MMTC) for Taiwan Mandarin (Tseng, 2013). We experimented with various subcorpora and across languages as illustrated in table 1 and with different potentials *base units*: syllables and letters for French; Characters, Pinyin (with and without tone) for Mandarin and only letters for English.

Corpus	# Tokens	# pseudo-Utterances
CID	125 619	13 134
MTR	42 016	6 425
MTX	36 923	5 830
DVD	64 023	7 989
DECODA (part)	580298	88 982
French	851202	122 360
MCDC	316 422	61 000
MTCC	122 200	26 000
MMTC	34 500	8 300
Mandarin	472 000	95 000
SWBD (English)	2 967 028	391 592

Table 1: Corpora used in the study

Some of those corpora are truly comparable while it is more debatable for others. MTR + MTX on French and MTCC for Mandarin are perfectly comparable since they have been recorded using the same protocol. CID for French and MCDC + MMTC are also very similar by nature. English Switchboard is perhaps a bit different in principle but in practice, it shares most of the features present in the previous corpora. The less similar of the set is French DECODA since it is recorded in a specific context (call center of Paris public transportation enquiries number). However, we add criterion during the extraction to try to avoid too many corpus specificities in a given language. Overall, all those corpora are truly conversational ones exhibiting the usual range of phenomena involved in fiercely spontaneous and interactional speech data. For all these corpora, the transcripts have been force-aligned at the word level.
Concerning the transcription, a standard orthographic transcription had been adopted for thes corpora. The spokenb particles do have standardized written forms in French (*euh, mh,...*) and English (*uh, um, mh...*). In the Taiwan Mandarin corpora, discourse particles, discourse markers, and fillers were transcribed with capital letters to distinguish themselves from foreign words such as English. Fillers are transcribed according to their phonetic forms. For instance, *UHN* is equivalent to *uhn* in English; MHM is something that is frequently observed in Mandarin, but not in English. In particular, multi-syllabic fillers are transcribed in one single unit, separated by H, e.g. *UHNHN*. See (Tseng, 2013) for more details.

4. Methodology

4.1. Segmentation

We use non-supervised machine learning algorithms (based on Branching Entropy, already applied to written Mandarin (Magistry and Sagot, 2012; Magistry, 2013)) for segmenting our sequence of characters coming from the conversational transcripts into our *base units* (spoken tokens). There are currently better methods for segmentation, especially for Chinese Word Segementation, but they require extremely large corpus that are not available for spoken language. Moreover, we were interesting in using the very same methodology on Mandarin, French and English with the idea in mind that the data set segmented in this same way across the languages could exhibit less divergence than being biased by the written form tradition of each language.
More precisely we use Eleve[2] *(Extraction de LExique par Variation d'Entropie - Lexicon extraction based on the variation of entropy)* toolkit. This method is helpful for our study because it allows us to get units grounded on the same principles and therefore not being biased by written processing techniques or conventions employed in different languages. Such an approach results in a new starting point for the type of lexical experiments we will perform later. An illustration of new units for French and English created by our approach are illustrated by Table 2. A benefit of such an approach is that we do not have to define what an individual word or multi-word expression is. We have done our experiments both with traditional segmentation (space-based) and with the output of unsupervised segmentation (in which, for example, *'you-know'*, turned out to be a unit). For a related work see (Dobrovoljc, 2017) which compare different association measures applied to discourse marker items.

While our unsupervised segmentation is very interesting to gather functional multi-words expressions into one unit as a result of the segmentation, it also presents some issues. For example, in French and English, it tends to split bound morphemes such as plural and gender marks as well as some verbal inflections. However, for our purpose of studying DM this feature should not be an issue.

[2] `https://github.com/kodexlab/eleve`

French	English	Mandarin
tu-vois	you-know	
mh-mh	uh-huh	MHMHM
ah-ouais	oh-yeah	對 A
c-est-vrai	that-s-right,that-s-true	
et-euh , donc-euh	and-uh, and-um	
et-puis, mais bon	and-then	
comme-ça	like-that	
dans-le, sur-le	in-the	
il-y-a	there-is	

Table 2: Examples of word like units created at segmentation stage *('-' in the units correspond to spaces in a traditional transcription)*

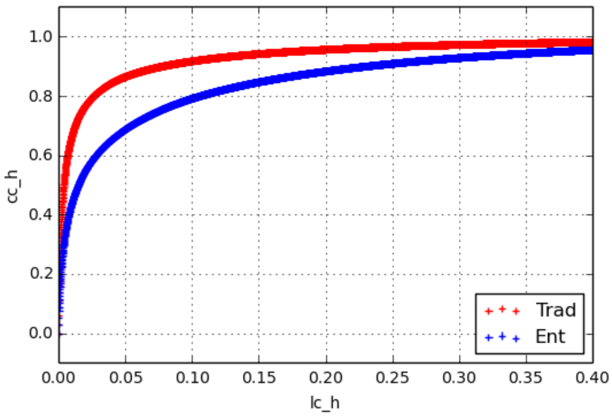

Figure 1: Comparative lexical growth (French) between traditional segmentation and Branching-entropy (x-axis : Coverage of the lexicon ; y-axis: Coverage of the corpus) segmentation

The unsupervised segmentation step provides a segmented corpus and a derived lexicon. In figure 1), we illustrate the lexicon coverage vs. corpus coverage of traditional vs. unsupervised segmentation.

In these corpora, we approximate the notion of utterance by using *Inter-Pausal Units* defined by continuous stretches of speech in between pauses of at least 200 milliseconds. Therefore, both our *lexical units* and our utterances are objective as possible, only relying on speech timoing and on the transcript.

4.2. Quantitative measures

Scores We argue that conversationally speaking, words distribution -Discourse Markers in particular- varies significantly depending on the type of utterances they occur in. A first relevant method being easy to apply in the study of conversation consists in separating the shortest sentences from the longer ones. Besides, it is a known fact that DMs can be found at specific positions in utterances (initial, median, final) with the initial and final ones being the most frequent (Aijmer, 2013; Filippi-Deswelle, 1998; Fraser, 1998; Muller, 2005; Stali, 2015; Stali, 2016). We propose to cross the two parameters mentioned above (type of utterance vs position in the utterance) to chart DMs.

Based on those two principles, we define a series of values aiming at characterizing quantitatively any form of the corpus (N: corpus size, S: number of tokens in short utterances, L: number of tokens in longer utterances, F_{all}: frequency of the token, F_{short}: frequency of the form in short utterances;F_{long}:frequency of the form in non-short utterances, F_{ini}:frequency of the form in initial position of longer utterances, F_{fin}:frequency of the form in final position of longer utterances

- $\frac{F_{all}}{N}$: relative frequency
- $\frac{F_{short}}{S}$: relative frequency of the form within short utterance forms
- $\frac{F_{long}}{L}$: relative frequency of the form within longer utterance forms
- $\frac{F_{short}}{F_{all}}$: tendency to occur in short utterances
- $\frac{F_{short}+F_{ini}+F_{fin}}{F_{all}}$: a sort of *"dm-hood"* of the form (tendency to occur in all canonical DM and interactional markers positions)
- $\frac{F_{ini}}{F_{long}}$: tendency to occur in initial position within longer utterances
- $\frac{F_{fin}}{F_{long}}$: tendency to occur in final position within longer utterances

We also use some of those scores to filter the set of items under consideration. More precisely we tested different thresholds for *relative frequency* and *dm-hood* scores. For French and Mandarin, we made sure that the relative frequency threshold was met for at least two-subcorpora in order to avoid domain-based items that could come from Maptask or DECODA corpora. This was both impossible and unnecessary to do on Switchboard corpus which is a lot larger and already more diverse thematically.

5. Experiments

In the context of this work, we were interested in comparing the clustering (and its implicit discourse marker characterization) in two approaches: traditional tokenisation and unsupervised segmentation. After segmenting the data sets and computing the scores as described in the previous sections we processed as follows. We filtered for relative frequency (threshold= 0.0005) and dm-hood (threshold= 0.3). Since we are at an exploratory stage of our work, those thresholds were chosen after inspection of results for a range of values for the both of them. We normalized all the resulting values, then applied PCA to the output and checked the *explained variance ratio* for deciding a number of principal components. The way DM are spread into the dimensions is illustrated for English DM in Figures 2 and 3 for traditional and unsupervised segmentation

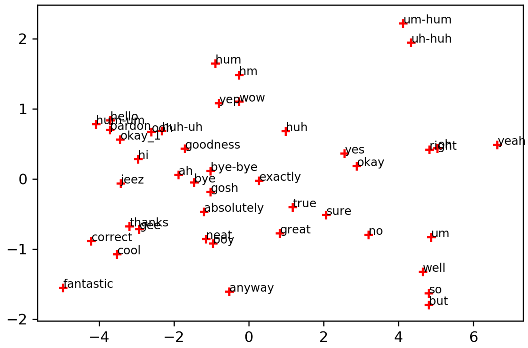

Figure 2: English DM plotted on the 2 principal components, based on traditional segmentation

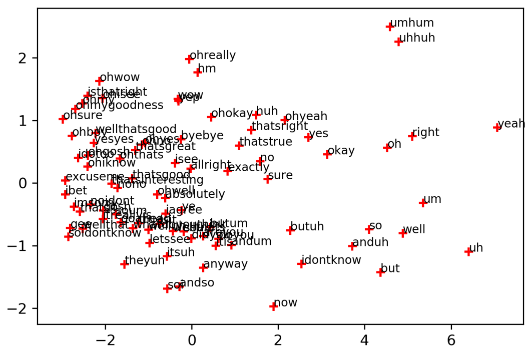

Figure 3: English DM plotted on the 2 principal components, based on unsupervised segmentation

respectively. Finally, after using the elbow techniques to determine an optimum number of clusters, we computed the clusters presented in Figures 4 and 5 for traditional and unsupervised segmentation respectively.

It is not straightforward to label the resulting clusters. However, it is possible to identify some known groups of markers in the clusters. We attempted to use the same color for similar clusters in both tradition and unsupervised results. Concerning traditional segmentation, the green and the yellow clusters host typical feedback items and a relative good match across languages. The division into two clusters is probably due to the fact that the items in the green cluster, in addition to be used frequently isolated as feedback items, may also occur in initial position (which is less the case for the 'yellow' items). Similar structure is observed for the unsupervised segmentation.

The 'blue' cluster corresponds to more evaluative and attitudinal items, at least for French and English. It is interesting to note that our very rough distributional measures are able to discriminate those items from the previous yellow and green clusters. We can see there is an adequate match between French and English but

ah,oh	oh	
ouais,oui	yeah,yes	
voilà	okay,right	
		EIN, EN, HON
mh	uh-huh,um-hum	MHM,MHMM,UHM...
d'accord,ok		
		HEIN,HEN,ON
alors,donc	so	然後, 所以
ben,bon	well	
et		
euh	um	NA
mais	but	可是
non	no	沒有
	sure	
		EI
		其實,就是,因為
cestca	correct	
exactement	absolutely,exactly	
hein		
hum		
putain	boy,goodness,gosh,jeez	
	anyway	
	bye,byebye,thanks	
	cool,fantastic,great,neat,gee	
	ah	HO
	true	
		就是說

Figure 4: Cluster (one per color) grounded on traditional segmentation

not so much for Mandarin.

Finally the red cluster includes at least two kinds of items: discourse connectives but also filled pauses and even interactional management items (French 'hein' in the unsupervised case). This is probably due to a lack of discrimination capacity for forms occurring within longer utterances at different places. For example, we know that *'hein'* tends to be more final but it is not enough to generate a specific cluster. Another explanation can be found in abandoned utterances. Those abandoned utterances typically end with a filled pause marker (French *'euh'*, English *'um,uh'*, Mandarin *'NEGE'*, *'NA'*. This (frequent) phenomenon therefore tends to make those items more distributionally similar to final particles like *hein*. Similarly, it may be rather surprising to see contrast connectives (*mais* / *but* / 可是) in those clusters. As mentioned above, this is probably due to unfinished utterances or utterance segmentation (based on pauses). However, in this cluster, there is a very satisfying match across the three languages.

6. Conclusion and Future Work

The exploration of DM spaces based on comparable corpora allowed us to show it remains possible to identify DM clusters, even through a cross-linguistic approach. The benefits of the unsupervised segmentation are not clear at this stage, specially for Mandarin

ah,oh		
ahoui,nonnon	no,ohno	
voilà	thatsright,thatstrue	對YA
ben		
	ohokay,okay,yes	
		EIN,EN,MHMM,ON,UNH
d'accord	right	對A
ouais,oui	yeah	
mh	uhhuh,unhum	MH
oh		
		HEIN,HEN
cestbon,cestvrai,cestça	sure	
eteuh,cesteuh,maiseuh,donceuh	andum,butuh,butum	
exactement	exactly	
tuvois	isee	
hum, maisbon,putain		
		HO,對不對
ahbon	ohreally,ohwow,wow	
	isthatright	
hm,hmhm	hm,huh	MHMHM, NHNHN, UHM
ok,toutàfait		
	yep	
alors,donc	so	
bon	well	
et	anduh	
euh	uh,um	NEGE,NA
hein		
mais	but	可是
non		
		EI,HON,其實,因為,就是,然後

Figure 5: Cluster (one per color) grounded on unsupervised segmentation

data. However, the method and approach adopted tend to demonstrate that the traditional segmentation already benefits from adapted transcription convention which includes rules for grouping specific words together. However, we believe it might be interesting to dig further in how much can be achieved without too many supervisions and bias from written resources. In the future, our first objective is to deeper scrutinize the elements in the structure of the Mandarin utterances which prevents DMs to be better clustered correctly with French and Mandarin items.

7. Bibliographical References

Aijmer, K. (2013). *Understanding pragmatic markers: a variational pragmatic approach*. Edinburgh: Edinburgh University Press.

Blache, P., Bertrand, R., and Ferré, G. (2009). Creating and exploiting multimodal annotated corpora: the toma project. *Multimodal corpora*, pages 38–53.

Crible, L., Bolly, C. T., Degand, L., and Uygur-Distexhe, D. (2015). Mdma: un modèle pour l'identification et l'annotation des marqueurs discursifs "potentiels" en contexte. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (16).

Dobrovoljc, K. (2017). Multi-word discourse markers and their corpus-driven identification. *International Journal of Corpus Linguistics*, 22(4):551–582.

Fernandez-Vest, J. (2015). *Detachments for cohesion: toward an information grammar of oral languages*, volume 56. Walter de Gruyter.

Fernandez, M. J. (1994). Les particules énonciatives dans la construction du discours. *Linguistique nouvelle.*

Filippi-Deswelle, C. (1998). *La relation dite de concession - Etude de Though, Although, Even Though et Even If antéposés en anglais contemporain.* Paris: Université Paris 7.

Fischer, K. (2006). *Approaches to Discourse Particles.* Amsterdam: Elsevier.

Fraser, B. (1998). Contrastive discourse markers in english? *Discourse markers: Descriptions and theory*, pages 305–312.

Fraser, B. (1999). What are discourse markers? *Journal of Pragmatics*, (31):931–952.

Furko, P. B. (2009). *The pragmatic marker - discourse marker dichotomy reconsidered - the case of 'well' and 'of course'*. dea.lib.unideb.hu.

Garric, N. and Calas, F. (2007). *Introduction à la pragmatique.* Paris: Hachette Education.

Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 517–520. IEEE.

Gorisch, J., Astésano, C., Bard, E. G., Bigi, B., and Prévot, L. (2014). Aix map task corpus: The french multimodal corpus of task-oriented dialogue. In *LREC*, pages 2648–2652.

Lenk, U. (1998). *Marking discourse coherence - Functions of discourse markers in Spoken English.* Gunter Narr Verlag Tübingen.

Magistry, P. and Sagot, B. (2012). Unsupervized word segmentation: the case for Mandarin Chinese. In *Proceedings of the 50th Annual Meeting of the ACL*, pages 383–387.

Magistry, P. (2013). *Unsupervised Word Segmentation and Wordhood Assessment.* Ph.D. thesis, Paris Diderot; Inria.

Muller, S. (2005). *Discourse markers in native and non-native English discourse.* John Benjamins Publishing.

Prévot, L., Gorisch, J., and Bertrand, R. (2016). A cup of cofee: A large collection of feedback utterances provided with communicative function annotations. In *Proceedings of 10th Language Resources and Evaluation Conference*, Portoroz.

Roze, C. and Danlos, L. (2011). Traduction (automatique) des connecteurs de discours. *TALN 2011, Montpellier*, (18).

Roze, C., Danlos, L., and Muller, P. (2010). Lexconn: A french lexicon of discourse connectives. *Proceedings of Multidisciplinary Approaches to Discourse (MAD 2010), Moissac, France.*

Schiffrin, D. (1988). Discourse markers. *Language*, (64):633–637.

Schourup, L. (1985). *Common discourse particles in English conversation.* New York: Garland.

Stali, M. (2015). *A corpus driven study: the use of dis-*

course markers during Twitch's streams. Master's Degree at Université d'Avignon.

Stali, M. (2016). *Les marqueurs discursifs genre et du coup: une étude comparative de corpus.* Master's Degree at Laboratoire Parole et Langage, Aix-Marseille Université.

Tseng, S.-C. (2013). Lexical coverage in Taiwan Mandarin conversation. *International Journal of Computational Linguistics and Chinese Language Processing*, 1(18):1–18.