

A Quantitative Comparative Study of Prosodic and Discourse Units, the Case of French and Taiwan Mandarin

Laurent Prévot

Aix Marseille Université

CNRS, LPL

Aix-En-Provence, France

laurent.prevot@lpl-aix.fr

Shu-Chuan Tseng

Academia Sinica

Institute of Linguistics

Taipei, Taiwan

tsengsc@gate.sinica.edu.tw

Alvin Cheng-Hsien Chen

National Changhai

University of Education

English Department

Changhai City, Taiwan

alvinworks@gmail.com

Klim Peshkov

Aix Marseille Université

CNRS, LPL

Aix-En-Provence, France

klim.peshkov@lpl-aix.fr

Abstract

Studies of spontaneous conversational speech grounded on large and richly annotated corpora are still rare due to the scarcity of such resources. Comparative studies based on such resources are even more rarely found because of the extra-need of comparability in terms of content, genre and speaking style. The present paper presents our efforts for establishing such a dataset for two typologically diverse languages: French and Taiwan Mandarin. To the primary data, we added morpho-syntactic, chunking, prosodic and discourse annotation in order to be able to carry out quantitative comparative studies of the syntax-discourse-prosody interfaces. We introduced our work on the data creation itself as well as some preliminary results of the boundary alignment between prosodic and discourse units and how POS and chunks are distributed on these boundaries.

1 Introduction

Interest for the studies of discourse prosody interface has arisen in the last decade as illustrated by the vitality of the events and projects in this domain. However, while theoretical proposals and descriptive works are numerous, quantitative systematic studies are less widespread due to the cost of creating resources usable for such studies. Indeed, prosodic and discourse analysis are delicate matters requiring lower-level processing such as the

alignment with speech signal at syllable level (for prosody) or at least basic syntactic annotation (for discourse). Moreover, many of these studies are dealing with read or monologue speech. The extremely spontaneous nature of conversational speech renders the first levels of processing complicated. Previous works (Liu and Tseng, 2009; Chen, 2011; Bertrand et al., 2008; Blache et al., 2009; Afantenos et al., 2012) give us the opportunity to produce conversational resources of this kind. We then took advantage of a bilateral project for working on conversational speech in a quantitative fashion, and this for two typologically diverse languages: French and Taiwan Mandarin. We believe this combination of linguistic resources and skills for these two languages is a rather unique situation and allows for comparative quantitative experiments on high-level linguistic analysis such as discourse and prosody.

Our objective is to understand the commonalities and the differences between discourse prosody interface in these two languages. More precisely, we look at how prosodic units and discourse units are distributed onto each other.

In spirit, our work is closely related to the one of (Simon and Degand, 2009; Lacheret et al., 2010; Gerdes et al., 2012), however our focus here are the insights we can get from a comparative study. Moreover our dataset has a more conversational nature than the datasets studied in their work. About the data, (Gerdes et al., 2012) wanted to have an interesting spectrum of discourse genres and speak-

ing styles while we focused on conversations both for making possible the comparative studies and to make sure to have enough coherent instances in the perspective of statistical studies. Also, while (Lacheret et al., 2010) requires a purely intuitive approach, we used a more balanced approach combining explicit criteria from different language domains. Finally, our annotation experiments are largely produced either by automatic tools (trained on experts data) or by naive coders. This is a major difference with the studies listed above that are based on experts annotations since it allows us scale up in data size more easily.

The paper is structured as follows. We will start in section 2 by presenting how we built a comparable dataset from existing corpora. Then we will address in section 3 and 4 respectively the creation of prosodic and discourse units. Based on these new datasets, we will investigate the discourse prosody interface in a comparative and quantitative way (Section 5). Finally, in section 6 we will pay some attention at what is happening syntactically at various types of boundaries as defined in the preceding section.

2 Building comparable corpora

lge	dur(m)	syll	tokens	PU	DU
fr	89	23631	20233	6057	2130
tw	205	54615	37637	8563	5673

Table 1: Size of the data set

First of all, corpora from both languages were recorded in very similar conditions. There are both face-to-face interaction in an anechoic room and speech was recorded via headsets on separate channels. The original recordings are also very comparable in size. The raw figures of both datasets are presented in Table 1.¹ We had to decide which linguistic information and which part from the full corpora to include in our joint dataset. About the later point, we extracted narrative sequences from the French data that included also more interactive topic negotiation sequences. About the linguistic levels, our study concerned prosodic and discourse levels but

¹See sections 3 and 4 for Prosodic Units, Discourse Units and Abandoned DU definitions.

we wanted to be able to perform fine-grained study involving syntactic and phonetic aspects. We therefore agreed to include syllables, tokens and part-of-speech information in our data as can be seen in Table 2. As the POS tagsets are different in both lan-

Description	Tier Name	Tier Content
Syllable	Syllable	STRING-UTF8
Token	Word	STRING-UTF8
Part-Of-Speech	POS	STRING-UTF8
Prosodic Units	PU	PU
Discourse Units	DU	{ DU, ADU }

Table 2: Contents of the joint dataset

guages, we established a matching table to make the POS information mutually understandable (Table 3).

tw	fr	Category
N	N	Nouns (<i>N</i>)
Nh	P	Pronouns (<i>Pro</i>)
Ne	D	Determiners (<i>Det</i>)
V	V	Verbs (<i>V</i>)
T,I,FW	I	Particles, DM ² ... (<i>Part</i>)
D	R	Adverbs (<i>Adv</i>)
A	A	Adjectives (<i>Adj</i>)
P	S	Prepositions (<i>Prep</i>)

Table 3: Correspondence of the most frequent POS tags

2.1 Creation of the French dataset

The ORCHID.fr Dataset is a subset of the Corpus for Interactional Data (CID) (Bertrand et al., 2008) consisting of 1.5 hour of conversational speech produced by 3 female and 3 male speakers. The CID corpus is a collection of 8 hours of free conversation in French. All the speaker turn boundaries are time-aligned with the speech signal at phone level by using forced alignment techniques (Illina et al., 2004). Moreover, the corpus had been entirely POS-tagged (See (Blache et al., 2008) for a presentation of the probabilistic technique used). Finally, in the framework of the OTIM and ORCHID projects an annotation campaign for annotating prosodic phrasing and segmenting the corpus into discourse units had been ran. In the present project, we modified the

criteria for labeling discourse units according to the commonly defined operational guidelines for French and Taiwan Mandarin data processing.

2.2 Creation of the Taiwan Mandarin dataset

The ORCHID.tw Dataset is a subset of the Taiwan Mandarin Conversational Corpus (the TMC Corpus), consisting of 3.5 hours of conversational speech produced by 7 male and 9 female speakers (Tseng, 2013). The TMC Corpus is a collection of 42 hours of free, task-oriented and topic-specific conversations in Taiwan Mandarin. All the speaker turn boundaries as well as syllable boundaries were human-labeled in the ORCHID.tw Dataset. Boundaries of words and POS tags were automatically generated based on the syllable boundary information and the output of the automatic word segmentation and POS tagging system developed by the CKIP at Academia Sinica (Chen et al., 1996). Previously, the ORCHID.tw dataset has been annotated with boundaries of prosodic units as defined in (Liu and Tseng, 2009) and with boundaries of discourse units in (Chen, 2011). In the present ORCHID project, we modified the criteria for labeling discourse units according to the commonly defined operational guidelines for French and Taiwan Mandarin data processing. The definition for prosodic units remains unchanged.

3 Producing prosodic units

3.1 French data

The definition of prosodic units is adopted mainly from prosodic phonology (Selkirk, 1986; Nespor and Vogel, 1986) that proposed a universal hierarchy of prosodic constituents. At least two levels of phrasing above the word have been admitted in French: the lowest level of phonological phrases (Post, 2000) or accentual phrases (AP) (Jun and Fougeron, 2000) and the highest level of Intonational phrases (IPs). The accentual phrase is the domain of primary stress. This latter is realized on the final full syllable of a word with longer duration and higher intensity than non-final syllables, and associated with a melodic movement. The secondary stress, more variable and optional, is generally realized on the initial stressed syllable of the first lexical word. It is associated with a rise movement.

The Intonational Phrase contains one or more accentual phrases. It is marked by a major final rise or fall (intonation contour), a stronger final lengthening and can be followed by a pause (Hirst and Di Cristo, 1984; Jun and Fougeron, 2000). More recently, a few studies attempted to show the existence of an intermediate level of phrasing (intermediate phrase, ip) that would be realized with stronger prosodic cues than the ones associated with AP and weaker than those associated with IP (Michelas and D’Imperio, 2010).

For the French dataset, both phonetic and phonological criteria have been used to annotate the boundaries of prosodic units. Once primary and secondary stresses are identified, the main acoustic cues are: (1) specific melodic contour, (2) final lengthening, (3) pitch reset. Moreover, disfluencies were annotated separately and silent pauses have not been systematically associated with a boundary (Portes et al., 2011). In a previous study involving two experts, we have shown the reliability of annotation criteria for the higher level of constituency (IP) (see (Nesterenko et al., 2010)). In a second stage, we elaborated a guideline for transcribing prosodic units in French by naive annotators. They have to annotate 4 levels of prosodic break defined in terms of a ToBI-style annotation (ref) (0 = no break; 1 = AP break; 2 = ip break; 3 = IP break) in Praat (Boersma, 2002).

Based on this break annotation we created *Prosodic Units (PU)* that are basically resulting from considering any break of level 2 or 3 as boundaries for our PUs. The merging of breaks of level 2 and 3 has been made to match the annotation style of the Taiwan Mandarin data but also to improve the reliability of the data produced. Indeed, the inter-annotator agreement was overall higher when levels 2 and 3 are collapsed. Finally, we added breaks on pauses over 400ms. We computed a κ -score for our data set by taking each token as a decision point and counting the number of matching and non-matching boundaries across annotators. This method of calculation yielded a κ -score of 0.71 for our dataset which is a nice score for naive coders on prosodic phrasing task.

Cohen’s kappa (Cohen and others, 1960) (and see (Carletta, 1996; Artstein and Poesio, 2008) for further discussion) is a measure designed to measure

inter-coder agreement. It corrects the raw agreement by an estimation of the agreement by chance. The issue here is that it is a segmentation task, therefore we have to decide on what are the decision points. We are using the tokens as decision points rather than a fixed sample (as it is done in some annotation tools) because the French guidelines are using words as the base units for instructing where to put the boundaries. Agreement on no-boundary (0-0) is therefore an agreement for this decision task and there is no satisfying way to evaluate a kappa score if these agreements are left out. Other measures need to be introduced (Pevzner and Hearst, 2002; Fournier and Inkpen, 2012) if one wants to measure a different aspect of the segmentation agreement. However to be perfectly transparent with the annotation results, Figure 1 presents the contingency table for the Orchid’s style prosodic units (See also (Peshkov et al., 2012) for deeper evaluation of the annotation of the whole CID corpus).

A/B	(0-1)	(2-3)
(0-1)	12242	1987
(2-3)	581	5272

Figure 1: Contingency table for the French prosodic units

3.2 Taiwan Mandarin data

The definition of prosodic units is adopted mainly from that of *Intonation Unit* in the field of discourse analysis (Chafe, 1994; Tao, 1996), but emphasizing on the concept of prosodic phrasing, instead of a coherent intonation pattern. We are in the opinion that prosodic phrasing is definitely not purely linear and sequential, as language planning should work with a certain kind of structure and hierarchy, which expectedly result in different types of prosodic phrasing. Nevertheless, the design of a single layer of prosodic phrasing will provide segmentation boundaries for further distinguishing the types of prosodic units and it is easier to achieve a reasonable inter-labelers agreement. Boundaries of prosodic units were annotated based on four main cues perceived by the labelers: (1) pitch reset (a shift upward in overall pitch level), (2) lengthening (changes in duration), (3) alternation of speech rate (changes in rhythm), and (4) occurrences of

paralinguistic sounds (disjunction or disruption of utterances such as pauses, inhalation, and laughter). The annotation of prosodic units of the ORCHID.tw Dataset has been accomplished in an earlier project (Liu and Tseng, 2009). Three labelers were trained to annotate prosodic units on a subset of 150 speaker turns until a satisfactory consistency rate was achieved. The rest of the dataset was completed by the three labelers independently.

Although the French and Taiwan Mandarin datasets were annotated based on different theories, but the annotation criteria were comparable. To ensure the comparability of the criteria, a cross-language segmentation experiment was conducted on a small subset of our data by the authors of this paper. Each tried to annotate prosodic units in the other language. The annotation results conducted by the non-native labelers confirmed that the main cues used for segmenting the prosodic unit boundaries were in principle uniform, except for those caused by repairs and restarts.

4 Producing discourse units

Concerning discourse units, the annotation campaign also involved naive annotators that have segmented the whole corpus (half of it being cross annotated). This annotation was performed without listening to the signal but with timing information. It was performed with Praat (Boersma, 2002) but without including the signal window, only the time-aligned token tiers. The segmentation was performed by adopting a set of discourse segmentation guidelines, inspired from (Muller et al., 2012) and (Chen, 2011). We combined semantic criterion (*Vendler’s (Vendler, 1957) style eventualities identification and Xue’s proposition identification (Xue, 2008)*), discourse criterion (*presence of discourse markers*) and pragmatic criterion (*recognition of specific speech acts*) to perform the segmentation.

More practically the task consisted in first identifying a main predicate, and then all its complements and adjuncts as illustrated in (1) and (2). Mandarin spontaneous speech presents an additional challenge in the task of DU annotation for its lack of tense-marking verbal system. Our segmentation proceeds on the basis of the semantic bonding between predi-

cates identified (Givón, 1993). Additional cues such as discourse connectives articulating discourse units were also used. Finally, mainly because of the interactive dialogic phenomena (e.g question-answer pairs) we added a few pragmatic criterion for allowing short utterances (e.g *yeah*) or fragments (e.g *where?*) (Ginzburg et al., 2007) to be acceptable discourse units.

(1) French Discourse Units

[on y va avec des copains]_{du} [on avait pris le ferry en Normandie]_{du} [puisque j'avais un frère qui était en Normandie]_{du} [on traverse]_{du} [on avait passé une nuit épouvantable sur le ferry]_{du}
[we going there with friends]_{du} [we took the ferry in Normandy]_{du} [since I had a brother that was in Normandy]_{du} [we cross]_{du} [we spent a terrible night on the ferry]_{du}

(2) Taiwan Mandarin discourse units

[qishi ta jiang de na ge ren yinwei ta you kai guo hui]_{du} [ta hai you jiang]_{du} [keneng shi ye bu zhidao wei she me]_{du}
[in fact the one he mentioned had the meeting]_{du} [he said in addition]_{du} [probably (he) did not know why, either]_{du}

Manual discourse segmentation with our guidelines has proven to be reliable with κ -scores ranging between 0.74 and 0.85 for the French data and reaching 0.86 for the Taiwan Mandarin data.

Moreover we distinguished between several units in discourse: *discourse units* and *abandoned discourse units*.³ The later are units that are so incomplete that it is impossible to attribute them a discourse contribution. They are distinguished from *false starts* (that are included in the DU they contributed) by the fact that the material they introduced cannot be said to be taken up in the following discourse unit.

(3) French abandoned discourse units

[et euh mh donc t(u) avais si tu veux le sam- + le]_{adu} [pour savoir qui jouait tu (v)ois]_{du}
[and err mm so tu had if you want the sat-

³We actually had also a *parenthetical* category but it was not consistently annotated at the current stage and therefore this distinction was not included in this paper.

+ the]_{adu} *[in order to know who play you see]_{du}*

- (4) Taiwan Mandarin abandoned discourse units
 [danshi muqian]_{adu} [yinwei shezhiyu]_{adu} [wo you ting renjia jiang]_{du} [man kuazhang]_{du}
[but for the moment]_{adu} [because even though]_{adu} [I heard some people say]_{du} [it is incredible]_{du}

5 Discourse Prosody Interface

5.1 Size of units

From Table 1, we can deduce the size of our units presented in Table 4. The significantly smaller French PUs (up to 40% depending to the units taken to compare) might partially be attributed to the difference in the segmentation style and the extraction of the subsets. The Taiwan Mandarin dataset contains only very long speaker turns, thus reducing the number of shorter prosodic units which are more often produced in interactive conversational speech. For DUs in which guidelines are basically identical we get very similar DU size in terms of duration and number of syllables (roughly 15% difference), French units host more tokens (43%) and therefore included shorter words.

	dur (s)	# syll	#tokens	# PU
PU-fr	0.88	3.9	3.3	-
PU-tw	1.44	6.4	4.4	-
DU-fr	2.51	11.1	9.5	2.8
DU-tw	2.17	9.6	6.6	1.5

Table 4: Comparative size of the units produced

Moreover from Table 1 we can see, that the French dataset included a larger part of abandoned discourse units (11% for 6,5% in the Taiwan Mandarin dataset). This is in line with the more spontaneous style conversations already mentioned in the French dataset.

5.2 Association of prosodic and discourse units

We examined the different types of association between prosodic and discourse units by means of boundary alignment. We follow (Chen, 2011) classification that starts from discourse units and that dis-

tinguishes 8 situations resulting from combining two parameters: (i) the presence of a prosodic boundary within the discourse unit (inner boundary vs. no-inner-boundary) ; (ii) the match of discourse and prosodic unit at either left, right, both or none boundaries. Such a classification resulted in the distribution illustrated in Fig 2.

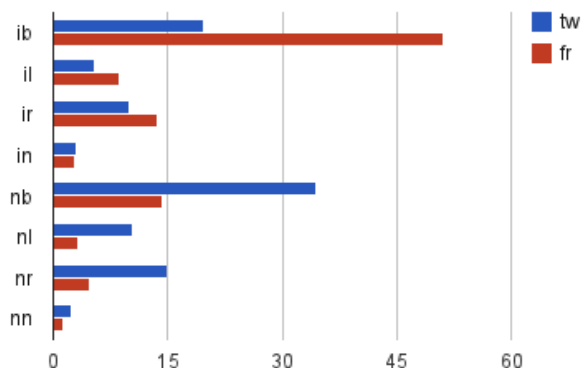


Figure 2: Distribution of PU/DU association types

In French data, perhaps because of the comparatively smaller prosodic units in the French data, discourse units host much more systematically several prosodic units. It is striking to see in figure 3 that more than half of the time and for both language discourse units are providing the starting and ending boundaries for the prosodic units. Overall, we see in figure 3 that once atomic and composite (in terms of PUs) DUs are collapsed their split in the alignment types are quite similar.

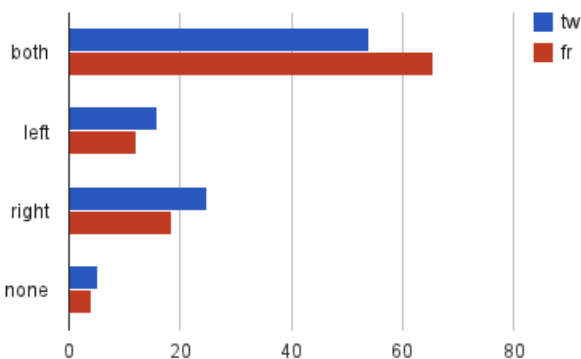


Figure 3: Distribution of PU/DU simplified association types

6 Syntactic categories at boundaries

Making use of the mapping table of POS information (Table 3) we established we are able to compare the distribution of POS at the boundaries. More precisely we looked at places where there was a match between PU-DU initial boundaries (Fig. 4) and PU-DU final boundaries (Fig. 5).

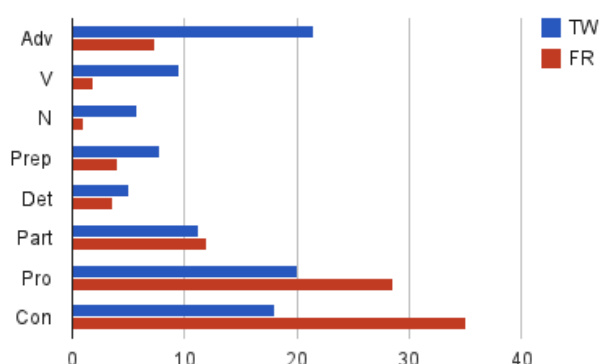


Figure 4: POS distribution at Initial matching boundaries

Interestingly, French units tend to begin more often with connectives and pronouns. In Taiwan Mandarin, the percentage of pronouns is lower and that of adverbs is higher. This may be due to fact that in conversation, sentences are often zero-subject or with the focus moved to sentence-initial positions.

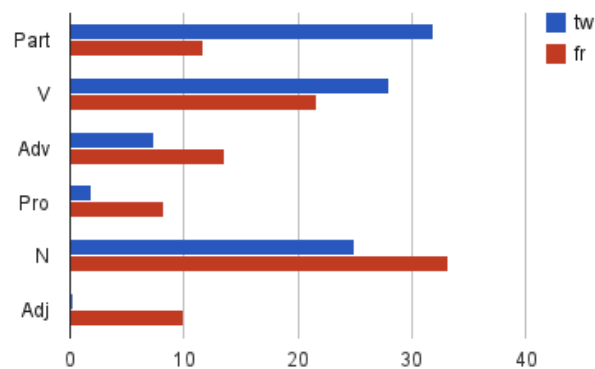


Figure 5: POS distribution at Final matching boundaries

For final matching boundaries, Taiwan Mandarin ends often with sentence-final particles, which is expected in conversation. Moreover, French ends more often at nouns than verbs, Taiwan Mandarin more verbs than nouns. Our preliminary studies on the word categories only provide information for the

boundary. More work on the sentence structure is required to conduct in-depth studies on language production.

7 Chunks as processing units

Chunks (Abney, 1991) can be seen as an intermediate level of syntactic processing. They are the basic structures built from the tags but do not deal with long dependencies or rich constituency. They are basically units centered on a syntactic head, a content word. As reminded by Abney, chunks can be related to ϕ -sentences (Gee and Grosjean, 1983) which have a more intonational nature. An idea defended in these early works is that chunks are indeed language processing units from a cognitive viewpoint. The break-up of experimental linguistics as renewed the interest for this hypothesis and is attempting to make it more precise (Blache, 2013) and relate to other empirical evidences such as eye-tracking (Blache and Rauzy, 2012).

With this idea in mind, we will investigate our prosodic and discourse units in terms of chunk size and constituency. The first basic hypothesis we are testing is if tokens are syntactic units and chunks more processing units, the structure of PUs and DUs in terms of tokens does not have to match across languages while it should in terms of chunks. More precisely, we expect a significant variation of PU/DU size across languages in terms number of tokens but not in terms of chunk size.

7.1 Creating chunks

VC	Verbal Chunk
NC	Nominal Chunk
AdvC	Adverbial Chunk
PC	Prepositional Chunk
IC	Intraactional Chunk
DisfError	Disfluencies or tagging errors
AdjC	Adjectival Chunk

Table 5: Chunks category created

From the chunking definition, we retain the importance of the head. We therefore designed simple rules using POS-tag patterns for creating the chunks listed in Table 5. This was done by looking at most

frequent patterns first. We processed in three different steps involving three different type of rules:

1. Propose for most frequent POS pattern a chunking rule (*e.g* $Pro\ Pro\ V \rightsquigarrow VC$; $Det\ N\ N \rightsquigarrow NC$)
2. Propose a set of rules aggregating tags and chunks into coherent chunks (*e.g* $Prep\ NC \rightsquigarrow PC$; $VC\ Part \rightsquigarrow VC$). This is done iteratively until stabilization of the number of sequences.
3. Simplification of the sequences by merging certain categories (*Det*, *Pro*) (or sequences of them) into some existing chunks (*e.g* $[Det|Pro]^+ VC \rightsquigarrow VC$) and simplifying some chunks sequences ($IC\ IC \rightsquigarrow IC$)

The two first types of rule are strongly language dependent while the third type is common to both languages.

Using pre-trained existing chunker was problematic. The rules used were defined to handle spontaneous spoken constructions. To our knowledge, existing chunkers are trained on written data which makes them impractical for our purposes. Moreover, in the rule-based design the rules are accessible to the linguists and this allow to compare them directly across languages rather than comparing chunking quality. Indeed, we are not interested in the chunks from an applicative perspective (such as named entity recognition) but as good approximation of semantic processing units. On the longer term, it could be however interesting to improve and evaluate and improve pre-trained chunking steps but this will require a large amount of manual work which we cannot afford for the time being.

7.2 Size in chunks

We then try to validate our hypothesis based on the chunks created and computed the size of PUs and DUs in terms of chunks (Table 6) and more precisely in terms of their length (in chunks) distribution (Figures 6 and 7).

Taiwan Mandarin and French size and size distribution exhibit however very different figures. About French PUs, it could be due to the sampling of the data (shorter PUs compared with the sampling of

lge	Size-PU	Size-DU
fr	1,48	3,69
tw	2,05	2,27

Table 6: Average size of units (in chunks)

long speaker turns data of Taiwan Mandarin) and the annotation criteria of PU. About the DUs, the distribution is also different but for this category we are more thinking at an issue with the tagging and chunking process. While we tried to keep the rules for producing the chunks coherent across the language, we might need either a more careful joint rules crafting or, perhaps a completely systematic chunking rules system. However, we do not have annotated chunks on this kind of data for training a supervised machine learning approach. Moreover, the dataset is significant but most likely not sufficient for unsupervised methods. In this context, crafting a simple rule-based system was appealing.

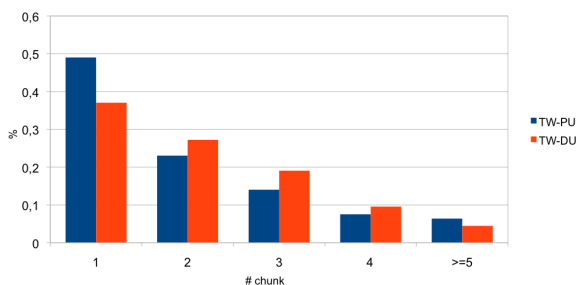


Figure 6: Comparison of Units size of the TW dataset

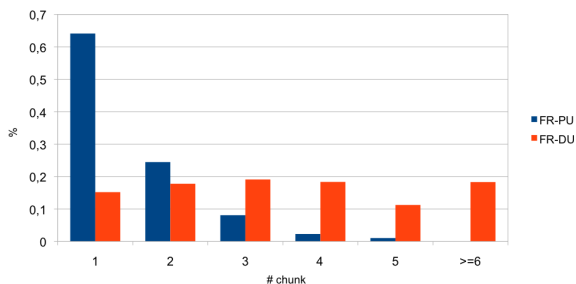


Figure 7: Comparison of Units size of the FR dataset

8 Conclusion and Future work

This work has shown that to create perfectly comparable corpora, one needs to start from joint design. However, this is a rare scenario and most of comparative datasets of richly annotated corpora will try to re-use at least part of their previous monolingual studies. Here we tried to make use of extremely similar resources for producing comparable corpora. We believe that although this data set could still be improved and benefited from an even more similar starting point, we have a unique resource for performing quantitative comparative studies of the kind initiated here. Equipped with this dataset, we are in position to conduct a series of deeper comparative studies. The chunking systems used in this paper are just a first attempt in this direction. Although the results for the chunk size are not conclusive for our hypothesis, we did get to know better the structures present in the units we are investigating and we would like to push further our exploration in this direction. We are currently looking at the distribution of the mono-,bi- and tri-chunks PUS and DUS sequences in order to get finer in the language comparison without going into a full syntactic analysis which is out of reach for this kind of data. In parallel, we will also attempt a shallower but more robust approach consisting in counting simply the number of content words in the units. This is even more basic than chunking but we would like to see whether it could be an interesting shortcut to the basic semantic structure of these units.

Acknowledgements

This work has been realized thanks to the support of the France-Taiwan ORCHID Program, under grant 100-2911-I-001-504 and the NSC project 100-2410-H-001-093 granted to the second author, as well as ANR *OTIM* BLAN08-2-349062 for initial work on the French data. We would like also to thank our colleagues for the help at various stage of the Data preparation, in particular Roxane Bertrand, Yi-Fen Liu, Robert Espesser, Stéphane Rauzy, Brigitte Bigi, and Philippe Blache.

References

- Steven Abney. 1991. Parsing by chunks. In Robert Berwick, Steven Abney, and Carol Tenny, editors, *Principle-Based Parsing*. Kluwer Academic Publishers, Dordrecht.
- Stergos Afantenos, Nicholas Asher, Farah Benamara, Myriam Bras, Cécile Fabre, Mai Ho-dac, Anne Le Draoulec, Philippe Muller, Marie-Paule Péry-Woodley, Laurent Prévot, et al. 2012. An empirical resource for discovering cognitive principles of discourse organisation: the annodis corpus. In *Proceedings of LREC*.
- Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.
- R. Bertrand, P. Blache, R. Espesser, G. Ferré, C. Meunier, B. Priego-Valverde, S. Rauzy, et al. 2008. Le cid-corpus of interactional data-annotation et exploitation multimodale de parole conversationnelle. *Traitement Automatique des Langues*, 49(3):1–30.
- Philippe Blache and Stéphane Rauzy. 2012. Robustness and processing difficulty models. a pilot study for eye-tracking data on the french treebank. In *24th International Conference on Computational Linguistics*.
- Philippe Blache, Stéphane Rauzy, et al. 2008. Influence de la qualité de l'étiquetage sur le chunking: une corrélation dépendant de la taille des chunks. *Actes, Traitement Automatique des Langues Naturelles*, pages 1–10.
- Philippe Blache, Roxane Bertrand, and Gaëlle Ferré. 2009. Creating and exploiting multimodal annotated corpora: the toma project. *Multimodal corpora*, pages 38–53.
- Philippe Blache. 2013. Chunks et activation: un modèle de facilitation du traitement linguistique. In *Proceedings of Traitement Automatique des Langues Naturelles*.
- P. Boersma. 2002. Praat, a system for doing phonetics by computer. *Glott international*, 5(9/10):341–345.
- Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational linguistics*, 22(2):249–254.
- Wallace Chafe. 1994. *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. University of Chicago Press.
- Keh-Jiann Chen, Chu-Ren Huang, Li-Ping Chang, and Hui-Li Hsu. 1996. Sinica corpus: Design methodology for balanced corpora. In *Proceedings of the Eleventh Pacific Asia Conference on Language, Information and Computation*, pages 167–176.
- A. C. Chen. 2011. *Prosodic phrasing in Mandarin conversational discourse: A computational-acoustic perspective*. Ph.D. thesis, Graduate Institute of Linguistics, National Taiwan University.
- Jacob Cohen et al. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.
- Chris Fournier and Diana Inkpen. 2012. Segmentation similarity and agreement. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 152–161, Montréal, Canada.
- James Paul Gee and François Grosjean. 1983. Performance structures: A psycholinguistic and linguistic appraisal. *Cognitive psychology*, 15(4):411–458.
- Kim Gerdes, Sylvain Kahane, Anne Lacheret, Arthur Truong, and Paola Pietrandrea. 2012. Intonosyntactic data structures: The rhapsodie treebank of spoken french. In *Proceedings of the Linguistic Annotation Workshop @ COLING*.
- Jonathan Ginzburg, Raquel Fernandez, Howard Gregory, and Shalom Lappin. 2007. Shards: Fragment resolution in dialogue.
- Talmy Givón. 1993. *English grammar: A function-based introduction*, volume 2. John Benjamins, Amsterdam.
- Daniel Hirst and Albert Di Cristo. 1984. French intonation: a parametric approach. *Die Neueren Sprachen*, 83(5):554–569.
- Irina Illina, Dominique Fohr, Odile Mella, Christophe Cerisara, et al. 2004. The automatic news transcription system: Ants some real time experiments. In *8th International Conference on Spoken Language Processing-ICSLP*, page 4.
- S.A. Jun and C. Fougeron. 2000. A phonological model of French intonation. *Intonation: Analysis, modelling and technology*, pages 209–242.
- Anne Lacheret, Nicolas Obin, and Mathieu Avanzi. 2010. Design and evaluation of shared prosodic annotation for spontaneous french speech: from expert knowledge to non-expert annotation. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 265–273. Association for Computational Linguistics.
- Y.-F. Liu and S.-C. Tseng. 2009. Linguistic patterns detected through a prosodic segmentation in spontaneous Taiwan Mandarin speech. In S.-C. Tseng, editor, *Linguistic Patterns in Spontaneous Speech*, number A25 in Monograph Series, pages 147–166. Institute of Linguistics, Academia Sinica.
- Amandine Michelas and Mariapaola D'Imperio. 2010. Accentual phrase boundaries and lexical access in french. In *Proceedings of Speech Prosody*.
- Philippe Muller, Marianne Vergez-Couret, Laurent Prévot, Nicholas Asher, Benamara Farah, Myriam

- Bras, Anne Le Draoulec, and Laure Vieu. 2012. Manuel d'annotation en relations de discours du projet annodis. Technical Report 21, CLLE-ERS, Toulouse University.
- Marina Nesper and Irene Vogel. 1986. *Prosodic phonology*. Dordrecht.
- Irina Nesterenko, Stephane Rauzy, and Roxane Bertrand. 2010. Prosody in a corpus of french spontaneous speech: perception, annotation and prosody~syntax interaction. In *Speech Prosody 2010-Fifth International Conference*.
- Klim Peshkov, Laurent Prévot, Roxane Bertrand, Stéphane Rauzy, and Philippe Blache. 2012. Quantitative experiments on prosodic and discourse units in the corpus of interactional data. In *Proceedings of SemDial 2012: The 16th Workshop on the Semantics and Pragmatics of Dialogue*.
- L. Pevzner and M. A. Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.
- Cristel Portes, Roxane Bertrand, et al. 2011. Permanence et variation des unités prosodiques dans le discours et l'interaction. *Journal of French Language Studies*, 21(1).
- Brechtje Post. 2000. *Tonal and phrasal structures in French intonation*, volume 34. Thesus.
- Elisabeth Selkirk. 1986. *Phonology and syntax: The relation between sound and structure*. The MIT Press.
- Anne Catherine Simon and Liesbeth Degand. 2009. On identifying basic discourse units in speech: theoretical and empirical issues. *Discours. Revue de linguistique, psycholinguistique et informatique*, (4).
- Hongyin Tao. 1996. *Units in Mandarin conversation: Prosody, discourse, and grammar*, volume 5. John Benjamins Publishing Company.
- S.-C. Tseng. 2013. Lexical coverage in taiwan mandarin conversation. *International Journal of Computational Linguistics and Chinese Language Processing*, 1(18):1–18.
- Zeno Vendler. 1957. Verbs and times. *The philosophical review*, pages 143–160.
- Nianwen Xue. 2008. Labeling chinese predicates with semantic roles. *Computational Linguistics*, 34(2):225–255.