Should we use movie subtitles to study linguistic patterns of conversational speech? A study based on French, English and Taiwan Mandarin

Laurent Prevot, Pierre Magistry, Pierre Lison

Should we use movie subtitles to study linguistic patterns of conversational speech? A study based on French, English and Taiwan Mandarin.

Laurent Prévot[1,2]*, Pierre Magistry*[3] *and Pierre Lison*[4]

[1] Aix Marseille Université, CNRS, LPL, Aix-en-Provence, France
[2] Institut Universitaire de France, Paris, France
[3] Aix Marseille Université, CNRS, IRASIA, Aix-en-Provence, France
[4] Norwegian Computing Center, Oslo, Norway

laurent.prevot@univ-amu.fr, pierre.magistry@univ-amu.fr, pierre.lison@nr.no

Abstract

Linguistic research benefits from the wide range of resources and software tools developed for natural language processing (NLP) tasks. However, NLP has a strong historical bias towards written language, thereby making these resources and tools often inadequate to address research questions related to the linguistic patterns of spontaneous speech. In this preliminary study, we investigate whether corpora of *movie and TV subtitles* can be employed to estimate data-driven NLP models adapted to conversational speech. In particular, the presented work explore lexical and syntactic distributional aspects across three genres (*conversational*, *written* and *subtitles*) and three languages (*French*, *English* and *Taiwan Mandarin*). Ongoing work focuses on comparing these three genres on the basis of deeper syntactic conversational patterns, using graph-based modelling and visualisation.

1. Introduction

Spoken social interactions constitute the most ubiquitous use of natural language for the vast majority of human beings. Achieving a more systematic understanding of the linguistic mechanisms at play behind these social interactions is thus of both theoretical and practical importance. In particular, the computational modelling of (human-human or human-computer) dialogues is becoming increasingly important for multiple applications, from virtual assistants to ambient computing devices, in-car voice control and human-robot interaction.

A major issue that hinders the development of data-intensive analysis of conversational speech is the relative lack of corpora and resources for this specific genre. Computational linguistics and in particular syntactic parsing have largely benefited from the creation of larger and richer *tree banks*. The first treebanks were build for written language, general news text or balanced written genres. As a consequence, the core building blocks of NLP pipelines such as tokenisers, taggers and parsers were designed, and later statistically trained, for such written data sets. It is from these generation of tools that all the NLP standards resources have been created resulting in a substantial bias for written forms and structures in NLP. This initial bias has worsened in recent years with the emergence of neural NLP models that often require very large amounts of training data to estimate their parameters. The only source of data that is considerable enough to provide many giga-words corpora is the web which is a again composed of various written genres (encyclopedic, news,...). Most of the transcribed conversational corpora have a size of roughly $10^4 - 10^6$ tokens whereas state-of-the-art NLP models are often trained on datasets containing $10^8 - 10^9$ tokens or more.

To address this issue, we explore the possibility of exploiting another source of data, namely *subtitles* from movies and TV series. Subtitles can be found in huge quantities and are available as open source data sets in various (raw or preprocessed) formats [1]. For the three languages we consider in this study (English, French, Taiwan Mandarin), the available data contains around $10^7 - 10^9$ tokens, with arguably a much more spoken and conversational genre than we can be found in traditional treebanks. We seek here to compare linguistic structures in these three languages across three genres: traditional written corpora, conversational spoken corpora and movies subtitles.

2. Related Work

Linguistics operated in the last decade an empirical turn in which corpora and annotated data sets of all kinds are the main sources of evidence [5]. Concerning specifically conversational spoken data, three disciplines need to be bridged: (i) *natural language processing* which treats language in a systematic way either based on pre-established rules or through statistical learning, (ii) *linguistics* that has applied, on this kind of data, rather descriptive and qualitative methods such as conversational analysis and more recently interactional linguistics [6][1] and (iii) *speech processing* that typically relies on shallow models learned over large (spoken or written) datasets that captures the most probable sequence of words and subsequently apply this knowledge for various speech tasks. Linguistic knowledge of conversational data is therefore scattered between rich descriptions of specific examples and shallow knowledge that often ignores the

[1] See however [7] for an recent effort of formalization.

	Written	Spoken	OpenSubtitles
French	GRACE	LPL [2]	sub fr
English	BROWN	SWITCHBOARD [3]	sub en
TW Mandarin	ASBC Sighan	MCDC22 [4]	sub zh

Table 1: *Datasets used in our comparison*

specific nature of conversational data.

2.1. Spoken language linguistics

While linguistics always had *speech* on its agenda, it must be noted that spontaneous spoken language structures received for a long time little attention from the linguistic community. One interesting exception is the spoken syntax movement [8] and some studies of spoken French [9]. While generally rejecting the idea of a different syntax for spoken languages (than for their written counterparts), this body of work highlighted significant differences in the way the linguistic system is implemented. They singled out phenomena rather specific to spoken languages and revised what is supposedly known about theses languages. We can mention the quasi-absence of Subject-Verb-Object clauses with lexicalized subject and object in spoken French while this was taken to be the prototypical clause organization in this language [10]. On a shallower level, it had been observed that the *type-token ratio (TTR)* was higher in written language [11] than in spoken language. In other words, one can find a higher proportion of content words per clause in written genres [12]. Finally spoken data tend to include a higher proportion of *Pronouns, Verbs, Adverbs and Interjections* while written data tends to include a higher proportion of *Nouns, Adjectives, Prepositions and Determiners* [13, 11].

2.2. Cognitive models of language use

The second important source comes from cognitive models of language production. The specificity of spontaneous speech [14] and of conversational speech [15] have been investigated for several decades. It lead in particular to empirically justified models of *disfluencies* [16] and of *conversational feedback*[17]. This work got integrated with earlier work on conversational data, with major findings on *turn-taking* structure [18] and *back-channels*[19]. All of these ingredients are crucial for understanding spoken language structures used in social interaction.

3. Data

We are using 9 data sets in our comparisons, crossing three languages and three genres, *written, conversations* and *subtitles*, as detailed in Table 1. The choice of the language is driven by the typological diversity and a practical reason of availability of transcribed conversational corpora.

The written corpora are balanced corpora that have been widely used in their communities, namely the French GRACE corpus, the Brown corpus and the part of the Academia Sinica Balanced corpus used in SIGHan Bake-of experiments. Conversational corpora come from different sources, respectively [20, 3, 4] while the subtitles are coming from the OpenSubtitles collection of movie and TV subtitles [1]. The OpenSubtitles collection is derived from more than 3.7 million subtitles (22 billion tokens) spread over 60 languages. Each subtitle undergoes a series of preprocessing steps (sentence segmentation, tokenisation, correction of OCR errors and inclusion of meta-data). The subtitles are then aligned with one another (both across languages and within the same language), allowing them to be exploited to learn machine translation models or other cross-lingual NLP applications. Recent work on neural conversation models also showed that subtitles can be used to train dialogue agents [21, 22].

Similar processing pipelines were applied to the aforementioned corpora to enable meaningful comparisons across experiments. For these first comparisons using standard automatic analyses on large data sets, we only require tokenizers and taggers for the three languages. For English and French, we used SPACY [23], a state-of-the-art NLP package, while we used Zpar for the Mandarin data [24] since there is at the moment no model for Mandarin in SPACY. The experiments described in this paper only employed a fraction of the subtitles corpora since the size of the conversational corpora was on a different scale.

4. Comparing Lexical Structures

We can first compare the shape of the lexical distribution for the different (tokenised) corpora and genres.

In the three *rank vs. frequency* plots of Figures 1, 2 and 3, we observe a similar pattern, namely that the subtitles corpus behaves as an intermediate form between the written and the conversational case. As illustrated by the three plots, the most frequent forms of the conversation data (and to a lesser extent of the subtitles data) are even more distributionally dominant than in the written case. This observation is compatible with the observation of higher *Type Token Ratio* for written genres. Figure 4 illustrates the average distribution for the three languages.

A simple analysis of the most frequent words in the different corpora can shed further light on this difference. Most of the standard function words (determiners, auxilaries,...) do no exhibit major differences across conditions. However, we can identify some groups that do show very strong differences across the copora:

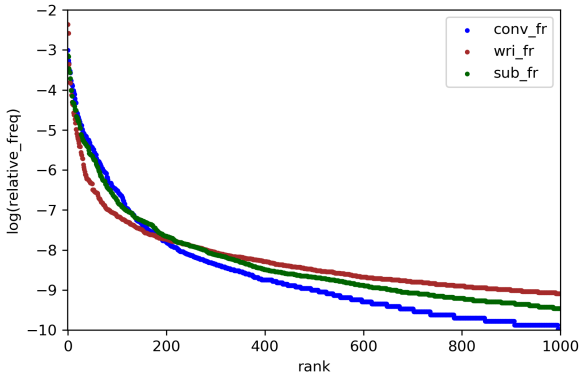

Figure 1: *Rank vs. Relative Frequ., French data*

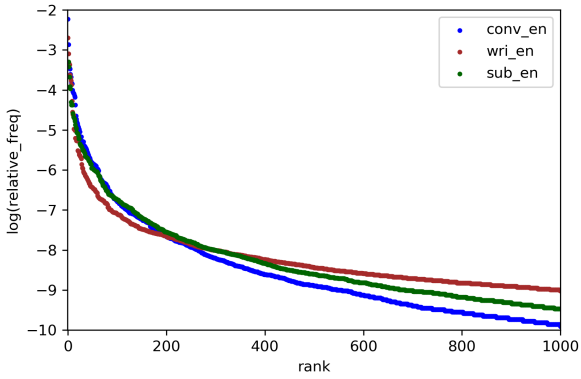

Figure 2: *Rank vs. Relative Frequ., English data*

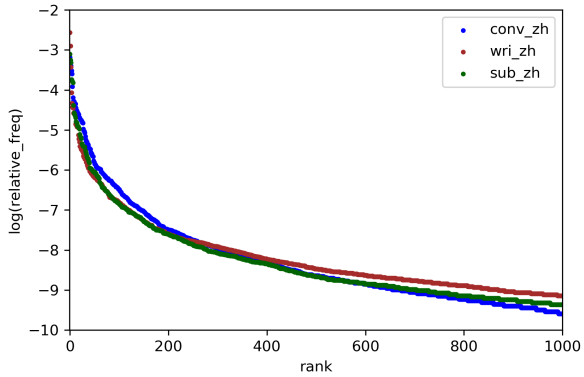

Figure 3: *Rank vs. Relative Frequ., Mandarin data*

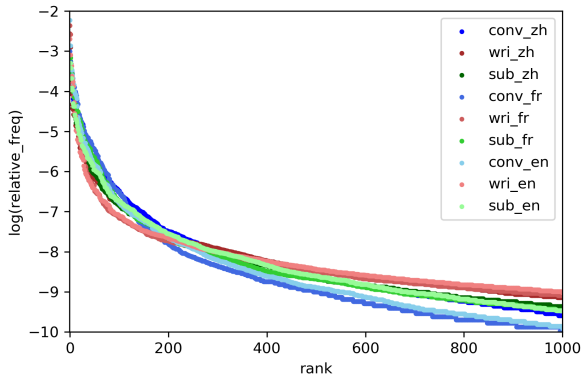

Figure 4: *Rank vs. Relative Frequ., All Three languages*

- Feedback back-channels markers (*French : ouais, euh, mh ; English : yeah, hm ; Mandarin: DUI-A, MH*), filled pause markers (*French : euh ; English : um, uh, huh, hum ; Mandarin: NEGE, NAGE*) are more frequent in the conversation corpus than in the subtitles one (and of course almost not represented in the written corpus). In fact, the filled pause (euh / uh-um) and the base back-channel *(French : mh, English : hm)* are simply absent from the subtitles while among the most frequent tokens in the conversations. We can note that the detailed transcription of Mandarin (conducted by linguists) ends up in a much more detailed set of back-channels and filled pause in the Mandarin conversation data than in the subtitles.
- Some of the feedback items (*French : , English : okay, yep, nope*), other interjections (*French : , English : wow, gosh*) and first and second personal pronouns (Fr : je, tu ; English : I, You) have similar relative frequencies in conversation and subtitle (while being relatively rare in written data).
- swear words and slang (*French : cool, vachement, pote, nana,... ; English : dude, bastard,...*), tend to be more present in subtitles than in conversations for English subset but better balanced in the French corpus[2] (and again near absent from the written corpus).

There are also some lexical content correlations related to the domains themselves. For example, content words on Switchboard are centered around everyday life, relation shop and the topic the participants are proposed to discussed about. Subtitles are derived from movies and TV episodes and exhibit therefore a stronger concentration of words related to this domain.

4.1. Comparing POS-tags distributions

The POS-tag distributions in the different corpora are illustrated in Figures 5,6 and 7.

The comparison of the simplified POS-tag distribution shows in the three languages considered that the

[2]This could be related to the distance between the speakers in Switchboard induced by the fact that they are talking over the phone with people the do not know. Interesting on this specific aspect that could make open-subtitle more similar to every day face-to-face interactions than Switchboard is.

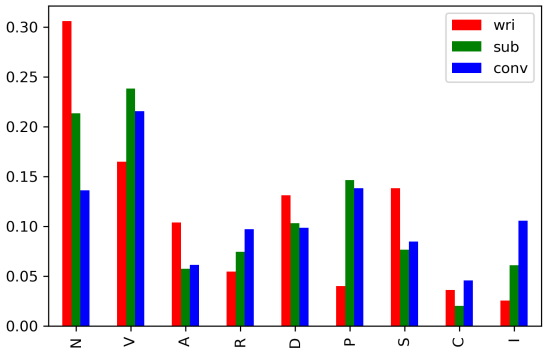

Figure 5: *POS-tag normalized distribution (English)*

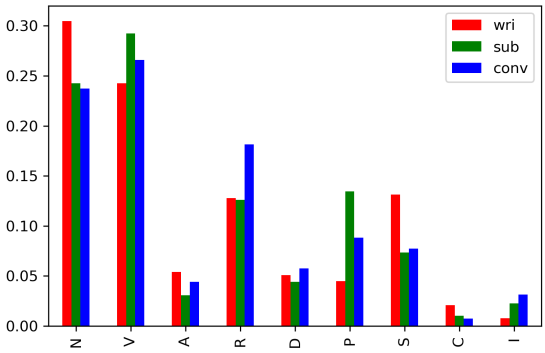

Figure 7: *POS-tag normalized distribution (Mandarin)*

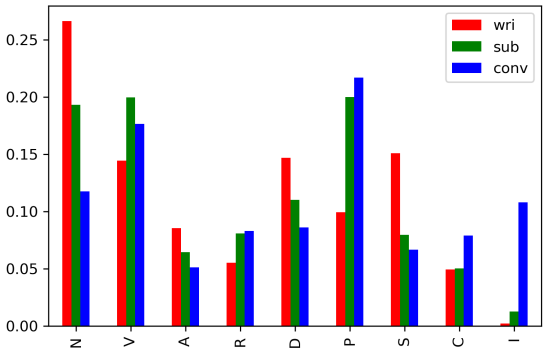

Figure 6: *POS-tag normalized distribution (French)*

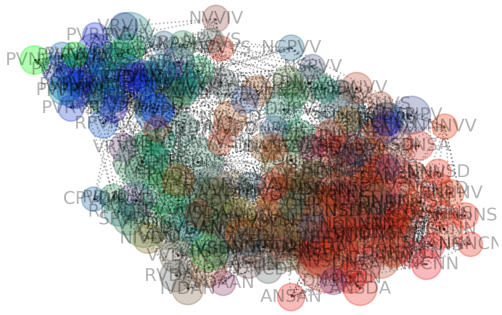

Figure 8: *Graphs of English POS-pentagrams, Red = Written, Blue = Conversation, Green = Subtitle, 200 patterns, Maximum edit distance = 2*

subtitles constitute an intermediate situation between the written and the conversational corpora. Verbs and Pronouns behave very similarly across the languages with *conversation* and *subtitles* having a very similar and higher verbs proportion than the written corpora. This inverted pattern is notably observed for Prepositions and to a lesser extent for Nouns.

Finally, there are some peculiarities that require further investigation, specially with regard to how interjections are handled with these standard tools.

5. Comparing POS-tag patterns

In order to conduct a deeper comparison of conversational syntactic patterns, we extracted POS-tag 5-grams from each corpora and built for each language a graph based on such POS-tag patterns. More precisely, the vertices are the POS-tag 5-grams (e.g PVPRV, PVDAN,...) and add an edge between two vertices only if the edit distance for the two patterns is very low. We experimented different weighting options (e.g differentiating between content vs function words tags) but settled down on a simple Levenshtein distance for the graphs presented in this paper. We then colored each graph according to the three genres.[3]

For illustrative purposes, we present below some examples of commonly observed POS sequences for each corpus type for English (1), French (2) and Taiwan Mandarin (3):

(1) a. *Conversational*: PVPRV, 'I guess we just hang', IPVPV : uh I am I think
b. *Subtitles*: PVDAN, 'He is an impressive man', SDNPV : 'of the problems I mean'
c. *Written*: DNSDN, 'the size of this city' ; NSNNN : 'result of city personnel policies'

(2) a. *Conversational*: PVPPV, 'on avait je je sais', *we had I I know*
b. *Subtitles*: PVRAV, 'on a bien sûr refusé', *we did of course refused*
c. *Written* : NSDAN, 'organisations d' un nou-

[3]The graph structure of Figure 8 is generated using the NETWORKX tool [25].

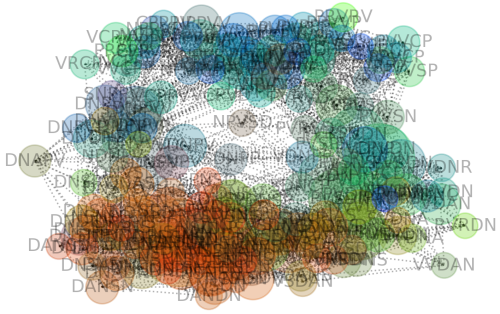

Figure 9: *Graphs of French POS-pentagrams, Red = Written, Blue = Conversation, Green = Subtitle, 200 patterns, Maximum edit distance = 2*

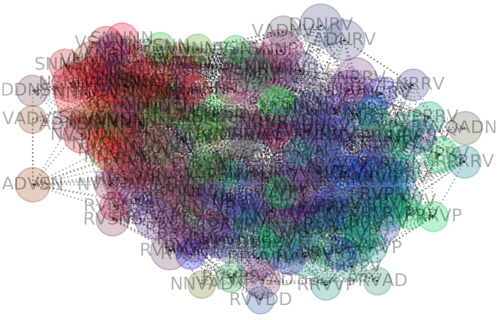

Figure 10: *Graphs of Mandarin POS-pentagrams, Red = Written, Blue = Conversation, Green = Subtitle, 200 patterns, Maximum edit distance = 2*

veau modèle' *organisations of a new kind*

(3) a. *Conversational*: VRVRV, 講 比較 清楚 一點 讓, *speak relatively clear a bit*
b. *Subtitles*: PRVRV, 我 還 要 多 練習, *I still have to practice more*
c. *Written*: NNSNN, 期間 院務 由 羅 副院長*period administrative duty by Luo vice-dean*

The graph-based visualisations produced illustrate that, even though the patterns are largely shared across the three genres, two clear areas (written and conversational) emerge, and this for all three languages. We observe also that subtitles (*more green*) pattern have a more complex distribution in the graph but, at least for French and English tend to create a specific smaller clusters.

6. Discussion and On-going work

The present paper compares three genres in order to understand whether *subtitles* can be fruitfully used for studying linguistic patterns of spontaneous speech. There are some obvious shortcomings since subtitles are avoiding *filled pauses* and the most colloquial (and frequent) back-channels feedback markers. The former was expected due to the scripted nature of movie and TV episodes. It is worth noting that even if filled pauses are present in the speech signal, they may not have been transcribed due to the crisp nature of subtitles, which need to adhere to strict time and length constraints [26]. The relative scarcity of feedback markers is slightly more surprising, since one could think these items are important to express the authentic feel of the interaction. However, putting these important differences aside, it is striking to see that *subtitle* corpora in the three languages are very similar to the *conversational* corpora in term of general lexical and POS-tag distributions.

Ongoing work focuses on systematizing the graph comparison along the lines presented in the previous section but also using existing graph comparison metrics on more extensive graphs (the graphs presented in the figures above were reduced to a "small" number of vertices for visualisation purposesà). This will allow a deeper evaluation of the relationships between the observed syntactic patterns. Another next step is to experiment with unsupervised segmentation of linguistic units (see e.g. [27] and even unsupervised tagging of these units, based on context clustering. The hypothesis we wish to investigate is whether the conversational units and structures of different languages are much closer than their written counterpart.

7. References

[1] P. Lison, J. Tiedemann, and M. Kouylekov, "Opensubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, 2018.

[2] L. Prévot, P. Magistry, and C.-R. Huang, "Un état des lieux du traitement automatique du chinois," *Faits de Langues*, no. 46, pp. 51–70, 2016.

[3] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, vol. 1. IEEE, 1992, pp. 517–520.

[4] S.-C. Tseng, "Lexical coverage in taiwan mandarin conversation," *International Journal of Computational Linguistics and Chinese Language Processing*, vol. 1, no. 18, pp. 1–18, 2013.

[5] M.-C. de Marneffe and C. Potts, "Developing linguistic theories using annotated corpora," *The Handbook of Linguistic Annotation*, 2015.

[6] E. Couper-Kuhlen and M. Selting, "Introducing interactional linguistics," *Studies in interactional linguistics*, vol. 122, 2001.

[7] J. Ginzburg, *The Interactive Stance: Meaning for Conversation*. Oxford University Press, 2012.

[8] C. Blanche-Benveniste, M. Bilger, C. Rouget, K. Van Den Eynde, P. Mertens, and D. Willems, "Le français parlé (études grammaticales)," *Sciences du langage*, 1990.

[9] K. Lambrecht, "Presentational cleft constructions in spoken French," *Clause combining in grammar and discourse*, pp. 135–179, 1988.

[10] ——, "On the status of SVO sentences in French discourse," *Coherence and grounding in discourse*, pp. 217–261, 1987.

[11] P. Fraisse and M. Breyton, "Comparaisons entre les langages oral et écrit," *L'Année psychologique*, vol. 59, no. 1, pp. 61–71, 1959.

[12] M. A. Halliday, "Differences between spoken and written language: Some implications for literacy teaching," in *Proc. 4th Australian reading conference, Adelaide, Australia, 1979*, vol. 2, 1979, pp. 37–52.

[13] H. Fairbanks, "The quantitative differentiation of samples of spoken language." *Psychological Monographs*, vol. 56, no. 2, p. 17, 1944.

[14] W. J. M. Levelt, *Speaking : from intention to articulation*. Cambridge, MA : The MIT Press, 1989.

[15] H. H. Clark and E. Schaefer, "Contributing to discourse," *Cognitive Science*, vol. 13, pp. 259–294, 1989.

[16] E. Shriberg, "Preliminaries to a theory of speech disfluencies," Ph.D. dissertation, University of California at Berkeley, 1994.

[17] A. Bangerter and H. H. Clark, "Navigating joint projects with dialogue," *Cognitive Science*, no. 27, pp. 195–225, 2003.

[18] E. A. Schegloff and H. Sacks, "Opening up closings," *Semiotica*, vol. 8, 1973.

[19] V. H. Yngve, "On getting a word in edgewise," in *Papers from the sixth regional meeting of the Chicago linguistic society*, 1970, pp. 567–578.

[20] L. Prévot, J. Gorisch, and R. Bertrand, "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, 2016.

[21] O. Vinyals and Q. Le, "A Neural Conversational Model," *CoRR*, vol. abs/1506.05869, 2015.

[22] P. Lison and S. Bibauw, "Not all dialogues are created equal: Instance weighting for neural conversational models," in *Proceedings of the 18th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2017)*. Saarbrücken, Germany: ACL, 2017, pp. 384–394.

[23] M. Honnibal and I. Montani, "spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing," *To appear*, 2017.

[24] Y. Zhang and S. Clark, "A fast decoder for joint word segmentation and pos-tagging using a single discriminative model," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2010, pp. 843–852.

[25] A. Hagberg, P. Swart, and D. S Chult, "Exploring network structure, dynamics, and function using NetworkX," in *In Proceedings of the 7th Python in Science Conference (SciPy)*, 2008.

[26] P. Lison and J. Tiedemann, "Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles," in *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, 2016.

[27] P. Magistry and B. Sagot, "Unsupervized word segmentation: the case for Mandarin Chinese," in *Proceedings of the 50th Annual Meeting of the ACL*, 2012, pp. 383–387.