Languages(s) of the SHUN-PAO, a Computational Linguistics account

Pierre Magistry

LANGUAGE(S) OF THE SHUN-PAO

A COMPUTATIONAL LINGUISTICS ACCOUNT

Pierre Magistry
Ph.D.
ENP-China, IrAsia, Aix-Marseille University
pierre.magistry@univ-amu.fr
Maison de la Recherche
29, av. Robert Schuman
13100 Aix-en-Provence
France

ABSTRACT / 摘要

This work is part of a broader project which requires adapting information extraction (IE) methods to written materials (mostly press articles) published in China between the mid 19th and the mid 20th centuries. This calls for a better understanding and description of the language(s) we can observe in our sources. More importantly, it is an unprecedented opportunity to provide a usage-based description of written languages as used in the press in Modern China. There is an abundant literature describing this pivotal era from different perspectives and disciplines related to language, including the history of language policies (Kaske, 2008), the socio-linguistic aspects (Weng, 2018) or historical linguistics (Coblin, 2000, Simmons, 2017). However what is presented in this article is, as far as I know, the first usage-based study to leverage a complete corpus of almost 80 years of a daily newspaper, the Shen-Pao（申報）, containing about 750 Millions sinograms to account for the actual practices and their evolution through time. In order to do so, I propose new Computational Linguistics methods and tools inspired by recent works in the field, especially Language Modeling and Contextual String Embeddings.

KEYWORDS / 關鍵字

Language change ; Language Modeling ; Lexical statistics ; Contextual Embeddings ; Modern China

1. INTRODUCTION

The work presented in this paper is conducted within the frame of a larger project, the “ENP-China, ERC Project”[1] which focuses on the transformation of the “elites”, their networks, and the exercise of power in Modern China, on a time span ranging from 1830 to 1949. This project intends to adopt and design a new methodology for “data-rich History”, by tapping into large collection of textual sources from that period. Our corpus includes daily news and relies on advanced Natural Language Processing (NLP) and Information Retrieval (IR).

This paper focuses on a descriptive study of one of our major sources, the Shun-Pao (申報, *shenbao*). It is the main Shanghai-based newspaper published during that period, whose publication spans from 1872 to 1949. Such study was required before going any further in the design of NLP tools for data extraction, but more importantly, it is an opportunity to provide an unprecedented data-driven account of language practices at that time.

1 https://enepchina.hypotheses.org/

The contribution also lies in the methodology used to face this large corpus of about 750M of sinograms (字), which is very coherent in terms of source (a single publication) but very heterogeneous in terms of time and language. I had to develop state-of-the-art NLP algorithms along with a software to navigate through the corpus and build a clearer yet global picture of the corpus, in order to provide a usage-based description of its language(s).

The beginning of the 20th century is a turning point in the socio-linguistic situation in China. It is marked by a flourishing diversity of cultural movements and political reforms that target language, and result in the birth of the National Language (國語 *guoyu*).

The very first line of our corpus, published on April 30th , 1872, is clearly in Classical Chinese: 「今天下可傳之事甚多矣」. The very end in contrast is very similar to Modern Chinese 「至大華戲院時，雨更大，大同學生秧歌隊也都在雨中淋着。」. What happened in between is what I intend to shed light on, with the constraint of adopting an approach solely based on the content of the corpus. Knowledge of the historical background marked by the political turmoil of the end of the Qing Dynasty, the establishment of the Republic, the cultural movements and the World Wars is here to motivate this study in the first place. This knowledge is also to be confronted to the findings, but was not used when defining the experiments presented in this paper.

The next section will provide some references and pointers to describe the historical context of the object of study and the challenges it raises for linguistics and NLP. I will then turn to a set of experiments designed to capture the main properties of the Corpus through time. The first one is focused on the use of punctuation marks, which is subject to evolution with two competing options: the traditional 句讀 jùdòu, which was used to annotate Classical Chinese for the convenience of learners and a Western inspired set of punctuation marks. The second experiment approaches the whole corpus on a yearly basis, with Language Modeling (LM) methods to enable us to run hierarchical clustering to spot different periods, which are internally homogeneous but distinct from each others. Theses periods are then considered as sub-corpora to build distributional spaces based on recent deep learning advances in NLP. The distributional spaces are explored through case-studies of sets of related words.

2. BACKGROUND OF THIS STUDY

The main contribution of this paper is to propose new computational methods to obtain a clearer picture of the Shun-Pao corpus. But before jumping into the computational details, this section provides some contextual information about the corpus, the socio-linguistic context of the time, and the challenges it represents for Computational Linguistics and NLP.

2.2 The Shun-pao and Chinese Languages at the turn of the 20th century

The Shun-Pao was established in 1872, targeting a well educated readership, who had received classical Chinese education. As early as 1876, it was experimented to add a supplement in vernacular language (白話 báihuà). However it did not met a public at that moment and its publication was stopped after a few months. Nevertheless, vernacular writings eventually became the language of the main publication. The journal was based in Shanghai but its readership extended to the rest of China, selling up to 30,000 copies a day.

A good overview of the language reforms, with a focus on education is provided by (Kaske, 2008). Reforms for a move toward vernacular languages started at the end of the Qing dynasty. Before the revolution occurred, a need for educational reforms was felt after the defeat against Japan in 1894. The language used for education was an important component in the debates on how to build a stronger country. The system of imperial examination was abolished in 1905. After the end of the Qing dynasty in 1911, this move towards vernacular languages continued, with new focus on mass literacy and nation building through the unification of language. Although the question on which language was to become the "National Language" (國語 guoyu) was still open. The complexity of the debates is on multiple layers, the classical Chinese/vernacular languages opposition does not stress the diversity of vernaculars, as vernacular literature

of that time includes many different regional languages. Even inside Mandarin, there was a strong opposition between Northern and Southern varieties (Pekinese and Nankinese). Account from historical linguistics (Coblin 2000, Simmons, 2017) attest that even though Pekin had been the capital of the Empire for a long time, it is only in the second half of the 19th century that the shift toward the Northern variety is well attested, and the Southern one kept a large part of its prestige (notably for being closer to the classical Chinese, and later because the northern variety was associated with the last, fallen dynasty). In the beginning of 1913 were held the meetings of the "conference for the unification of reading pronunciations" (讀音統一會), reports about these meeting tend to confirm the unsettled situation at that time (see Kaske, chapter 6). In 1919 the Ministry of Education founded the "Preparatory Committee for the Unification of the National Language", which Kaske describes as "the first permanent language planning institution" and the actual beginning of the institutionalization of what would eventually become Modern Standard Chinese (MSC). Education policies from this committee are however unlikely to have affected the writers of the Shun-pao (considering their age).

Simmons (2017) claims that "*Běijīng was not firmly established as the norm until the People's Republic of China definitively declared the city's dialect as standard in the 1950s.*" and describes the koiné of the early Republican period as 藍青官話 lán-qīng guānhuà 'blue-green (impure) mandarin' to denote its composite nature. Going even further, in 1967, Paul Kratochvíl was describing MSC or Pǔtōnghuà as "*the language used today by educated speakers of North Chinese dialects (primarily Peking Dialect) which most speakers of Chinese consider as the correct form of oral communication and* ***which is also the basis of the slowly emerging modern written norm***". In the introduction he stresses the difficulty to conduct a descriptive work when the question of establishing the norm is still open. I shall note however, that not everybody shares this view of a very late establishment of Pekinese as the norm. For exemple (Huang, 2014) in a paper focusing on some differences between Mandarin in China and in Taiwan states that "*a variant of Northern Mandarin Chinese, is designated as the common language about a hundred years ago*".

A full review of language policies and their actual effects is out of the scope of this article. I can only recomand the reading of Kaske and Simmons works, but it shall be clear that the Shun-Pao was written at a time when the national language was the object of vivid debates and cultural movements. We can expect to find different stages and competing options toward the establishment of the national language, all the way back to classical Chinese under the late Qing dynasty.

A sound methodology could be to focus on the history of the journal and its actors ("who owns it ?", "who writes it", "who reads it ?"), but in this work I focused on the language in itself, "what has actually been printed ?". This calls for a data-driven approach, the results shall be compared to the historical or socio-linguistic aspects in a future work. We will start by using computational methods to have the printed material "speak for itself".

2.2 Language Diversity and NLP

With the exception of very recent trends in computational socio-linguistics and NLP applied to historical linguistics, most of the works in NLP rely on a definition of Language (as the object to be studied or processed) which follows a very strong assumption from formal linguistics. That of an "homogeneous community of speaker-listeners". As Chomsky (1965, p.3) puts it :

"Linguistic theory is concerned primarily with the ***ideal speaker-listener****, in a* ***completely homogeneous speech community****, who knows its language* ***perfectly*** *and is unaffected by such grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interest, and* ***errors*** *(random or characteristic) in applying his knowledge of the language in actual performance. This seems to me to have been the position of the founders of modern general linguistics, and no cogent reason for modifying it has been offered"*

This approach of language is not really adapted to the situation of Modern China. Where the society is multilingual, and a national language yet to be defined and standardized (which makes the expressions "errors" and "perfect knowledge of the language" meaningless).

The more noticeable move away from this assumption in the NLP community is the stream of works focusing on computational methods and resources for related languages and language varieties, especially the

VarDial serie of workshops and evaluation campaigns. But in this field, only varieties of present-time Mandarin are getting attention. A pioneer work in this direction is the LIVAC Corpus maintained at the City University of Hong-Kong since 1995. The most recent at the time of writing was the evaluation campaign of VarDial 2019 (Zampieri et al., 2019) in which one shared task focuses on discriminating between Mandarin Chinese in a news corpus including data from China and Taiwan. It is noteworthy that the paper presenting the shared task claims that Mandarin "has been the official language of the government by convention for over a thousand years but has also become the common language both in spoken language and written text by constitution in the modern era, first by the Nationalists (ROC) after 1911, and then by the Communists (PRC) in 1949." without any consideration for the historical and socio-linguistic complexity.

To my knowledge, no annotated corpus or computational study of written language is available for our type of source on the pre-1949 period.

Beside an inquiry into the late history of what was about to become Modern Chinese, this work is also motivated by very practical needs inside the ENP-China project. In order to apply relevant analysis tools and models, we need a clear picture of the targeted data. At first sight, it seems that NLP tools trained for Modern Standard Chinese may work on the later part of our corpus, but not on the beginning. It is well established that discrepancy between training and testing data can be extremely harmful to NLP tools. For example, (Bamman et al., 2019) report a loss of 20 points in f-score in Named Entity Recognition (NER) when training on a news corpus and testing on literature. In our project, we will need to build our own training corpora, hence the need of a sound way to describe, split and sample the corpus.

Most campaigns from VarDial and other works on language identification frame the question into a classification problem on which one can apply supervised machine learning. This is not applicable in our case as we have no training data, and not even a clear picture of the different classes to be found. We must adopt a more descriptive approach. The tools I propose in the following sections are made to ease the exploration of this large corpus using clustering and unsupervised learning. I do not attempt any strict classification for the moment.

3. PROCESSING AND TOOLING OVERVIEW

The source material of this study is a full-text version of all the published issues of the Shun-Pao. It was provided to us a a collection of "plain/text" files, with some metadata (such as the date or issue number) inserted in the text. It was possible to extract these metadata with few regular expression. The article segmentation is not included but can be made (imperfectly) based on heuristic rules.

Once this basic corpus structure has been identified, it is rendered as XML and stored in an eXist-DB database. At this stage, the actual content is indexed by Lucene which is embedded in eXist-DB. This allows for simple lookup and occurrences statistics needed to provide n-grams counts and punctuation overview.

The rest of the experiments required more coding. I used Scala to query the eXist-DB, prepare the data and call external tools such as KenLM (Heafield et al. 2013), which is used for training and querying Language Models (LMs). Python is used for the Deep-Learning part (using Flair/pytorch) and R is used to build and share the clustering and other data visualization, thanks to R-Shiny.

In some cases, it was tempting to apply some text normalization as a preprocessing step to the digitized corpus. This is especially the case for sinograms with graphical variants which we may want to unify. For exemple, 没 and 沒 are both present in the text files, with a disturbing pattern which is likely to be a consequence of using different OCR systems and/or inconsistent manual correction of its output. On the other hand, in many other cases the existence of competing variants is worth studying, and a natural phenomenon at a time when writing was not strictly standardised. I will show below that with the methods I propose, we can observe the competition between 很 and 狠, which actually provide good cases to assess for the relevance of the computation. I thus decided to keep the text as provided, and to only rely on variant dictionaries to expand my queries.

4. NOTE ON THE FIGURES

All the figures presented in this paper (and more) are available online and interactive on the R-Shiny platform located at https://analytics.huma-num.fr/Pierre.Magistry/Shun-pao_DADH2019/. I invite the reader to consult this dedicated website while reading the paper to try different parameters on these experiments and get a more precise picture of this study.

5. PUNCTUATION

The first striking difference between the beginning and the end of the corpus, which can be spotted even without knowledge of Modern nor Classical Chinese, is the punctuation. It goes from being absent to a western style punctuation. Mullaney (2017) gives a fascinating account of this change. He describes a first stage that introduces traditional punctuation (句讀 jùdòu) as an annotation to ease the reading. This stage involved professional "punctuators" who were different from the authors of the articles. After that, under the pressure of language in contact at the script level (mixing Chinese and Latin characters) and with more westernized and multilingual education of the journalists, a shift to western style punctuation. The description given by Mullaney is consistent with what we can observe in the Shun-Pao corpus, but he did not give precise dates and statistics that we are now able to provide.

To do this, I simply extracted counts of different punctuation marks to compute their frequencies on a daily/monthly/yearly basis (in number of punctuation mark occurrences per characters). I tried a large set of marks including different styles of parenthesis, exclamation or interrogation marks, etc. But the whole story can be told with only three marks : the full stop 『。』, the comma 『，』 and the "enumeration comma" 『、』. The comma is the last one to appear in the corpus, and the function of the two other evolved through time.

The graphic obtained from these counts is presented on Figure 1. It shows a first period of experimentation with the full stop in roughly 1906~1911, it goes back to no punctuation until the 1920's. Then we see a competition and an inversion of the frequencies of the two commas during the 30's and 40's.

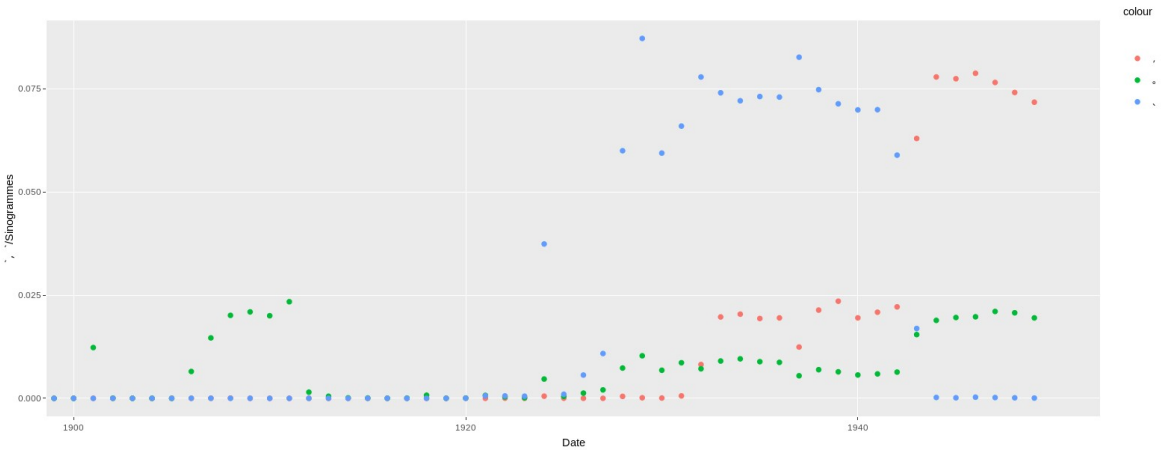


Figure 1: Frenquency of punctuation marks

Based on this global picture, we can now have a closer look at the original issues on specific dates. In 1909, we can see examples of classical punctuation as annotation beside the text. In this case it is worth

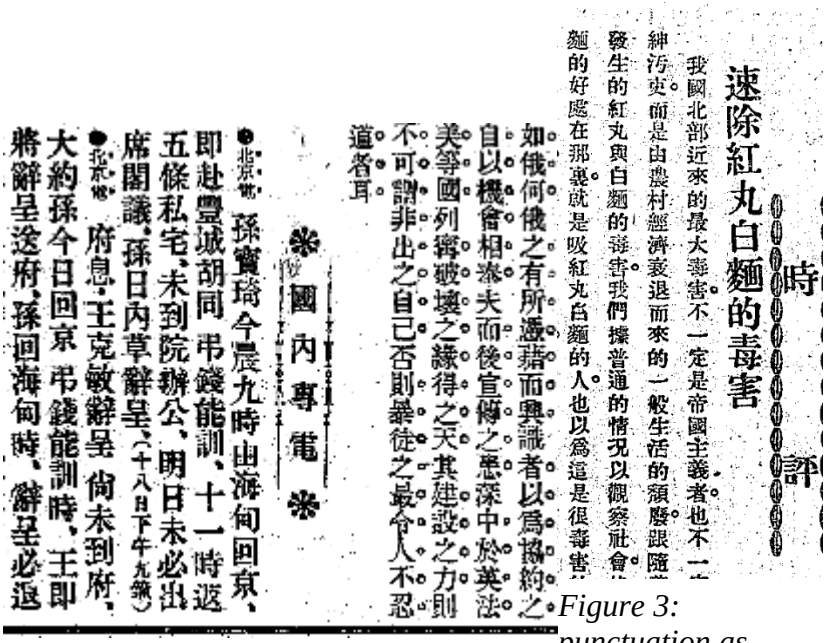
noting that the transcription to full-text alter the actual content as it renders the text with full stops inserted in the text. The beginning of the text in Figure 2 is actually transcribed as 「嗟夫悲哉。華僑之別祖國焉久矣。」. It is already a conversion to a more modern punctuation and it explains why the frequency of the full stop is still quite low in our figure.


論華僑之可憫

嗟夫。悲哉。華僑之別祖國焉久矣。隨其父兄經商於絕域。或因壯年其宗旨。皆無非欲少得金貲。還返精工即倣其神錨銖以積之分寸。得當則積貲數十百萬。以至千萬。仍不得不滯跡於異邦。不幸而不更無餘資可爲買田歸隱之計。又想蓴鱸之味。蓋其心皆有所不得已也。乃他人見我人數多而工價

Figure 2: Punctuation as annotation on the right of the text (1909)

In the first half of the 1920's, we can observe the coexistence of this punctuation as annotation style and in other cases full stops and traditional comma 『、』 inserted in the flow of the text. The two styles can even occur on the same page, as illustrated on Figure 3.


時評

速除紅丸白麵的毒害

我國北部近來的最大毒害。不一定是帝國主義者也不一定是紳汚吏。而是由農村經濟衰退而來的一般生活的頹廢。跟隨發生的紅丸與白麵的毒害。我們據普通的情況以觀察社會。麵的好處在那裏。就是吸紅丸白麵的人。也以爲這是很毒害。如俄何俄之有所遵藉而興。識者以爲協約之自以機會相乘。夫而後宣傳之惡。深中於英法美等國。列強破壞之隙。得之天。其建設之力。則不可謂非出之自己。否則暴徒之最令人不忍道者耳。

Figure 3: punctuation as annotation in the margin (1934)

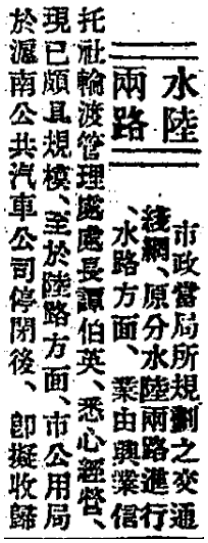
國內專電

●北京●孫寶琦今晨九時由海甸回京、即赴豐城胡同弔錢能訓、十一時返五條私宅、未到院辦公、明日未必出席閣議、孫日內草辭呈、(十八日下午九鐘)

●北京●府息、王克敏辭呈尚未到府、大約孫今日回京弔錢能訓時、王即將辭呈送府、孫回海甸時、辭呈必退

Figure 4: Two articles with different punctuation styles on the same page (1924)

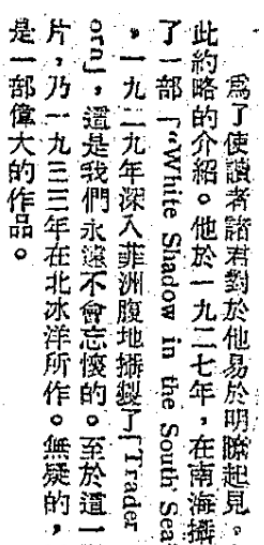
Then the western style comma rises to reach a first plateau between 1933 and 1942. During that period, We can observe that some articles like the edito keeps a more traditional style of full stops as annotations on the side of the text. Their use however is reversed as it is the presence, not the absence of a full stop that marks the end of a clause. The rest of the issues are mixture of articles with either no punctuation or different schemes inserted in the flow of the text (only 「、」, combination of 「、」 and 「。」 or combination of 「，」 and 「。」). As in Figures 5 and 6.



水陸兩路

市政當局所規劃之交通綫網、原分水陸兩路進行、水路方面、業由興業信托社輪渡管理處處長譚伯英、悉心經營、現已頗具規模、至於陸路方面、市公用局於滬南公共汽車公司停閉後、即擬收歸

Figure 5: (1934)



爲了使讀者諸君對於他易於明瞭起見。在此約略的介紹。他於一九二七年，在南海攝製了一部「"White Shadow in the South Sea"」。一九二九年深入菲洲腹地攝製了「Trader Horn」，這是我們永遠不會忘懷的。至於這一影片，乃一九三三年在北冰洋所作。無疑的，這是一部偉大的作品。

Figure 6: (1934)

Starting in 1943, the frequency of western 「，」 becomes higher that the frequency of 「、」, which ends up limited to its modern enumeration function.

In this Section, I showed that simple frequency statistics of punctuation marks can point to interesting dynamics which testify for the lively and creative aspects of publishing in Modern China. As written language is the object of major political debates, different competing writing and printing styles can appear on the very same page of the Shun-Pao.

This global approach to provide statistics based on daily/monthly/yearly counts can already exhibit different periods. It begins with no punctuation at all, then the traditional marks used as annotation to ease the reading are added on the side of the text to ease the reading around 1906. in the 20's, the punctuation is inserted in the flow of sinograms. It becomes more frequent and uses traditional marks. In the 30's, a more westernized punctuation appears but it is only after 1943 that it is widely adopted and turned into the norm in the journal.

It is worth noting that when going back to the source images for verification and to seek actual examples, I noticed that the OCR Is not always reliable. More importantly, the transcription of the punctuation as annotation is not always consistent. It is usually inserted into the text and often simply discarded. The graphic on Figure 1 may give the impression that punctuation practice stops between 1913 and 1924, but it is only its transcription which is missing. This does not challenge the whole story just depicted, but these inconsistencies (together with the genuine variation which comes from the competing styles) make it difficult to rely on the punctuation for further computation over the whole corpus. For that reason, I decided to discard the punctuation from the transcriptions in the subsequent experiments.

6. CLUSTERING WITH LANGUAGE MODELS

In this section I propose to spot different periods in the corpus based on the text using language modeling. A language model (LM) is made to estimate a probability distribution of sequences of tokens, here the tokens are sinograms. A LM will assign a probability to any sequence of tokens, which should be high if the sequence is similar to the language it models and low otherwise. An important measurement in language modeling it the *perplexity (PP)*. It gives a value of how well a model is able to account for a sample. If a sample is drawn from a distribution similar to the one of the model, the perplexity will be low. It will be high otherwise. One can think of the model being "perplexed" at the sight of a weird sample.

The main idea for this experiment is to use the perplexity as the basis to define a metric to apply hierarchical clustering of the different parts of the corpus. To have enough data to estimate a LM on each subcorpus but still have relatively fine grain clusters, I chose to split the corpus into one sub-corpus per year. I then use KenLM to estimate a LM for each year. With one sub-corpus and one LM per year, it is now possible to use perplexity of the LMs to define a distance measure between every two years of the Shun-Pao.

If we consider two years A and B of the Shun-Pao, and the corresponding language models LM_A and LM_B, we can compute the perplexity PP_{AB} of the LM_A in front of the text of B and the perplexity PP_{BA} of the LM_B in front of the text of A. The perplexity is not a symmetric measure (PP_{AB} can be different from PP_{BA}) so it cannot be used directly as a distance metric for clustering. To obtain an actual distance measure between two years A and B, we can simply sum the two perplexities and define $dist(A,B) = PP_{AB} + PP_{BA}$. I this case, dist(A,B) = dist(B,A) and we can use this value for a sound clustering.

Once the distance matrix is built, we can apply multidimensional scaling (MDS) to visualize the data on a two-dimensional plane and agglomerative clustering to define periods over the whole corpus. Considering the nature of the data, a stronger relation is to be expected between every two adjacent years, not as a consequence of language similarity but because of the topicality of current events at the time. For this reason, some agglomeration methods such as "single linkage" are expected to perform poorly, and I focus on other methods less sensitive to such issue like complete linkage or Ward's method (as implemented in R *hclust* function).

The resulting plot of MDS and clustering dendrograms are presented below. Both clustering methods seems to agree on relatively clear cuts after 1904, 1911 and 1937. The pre-1904 period is split either after 1894 (Ward) or 1892 (complete linkage) and another small disagreement occurs between 1921 (complete linkage) and 1926 (Ward). Overall the two clusterings seems consistent, with another difference which is that Ward consider 1904 as the more salient cut, followed by 1926 and 1911 where complete linkage stresses 1921 before 1904 and then 1911.

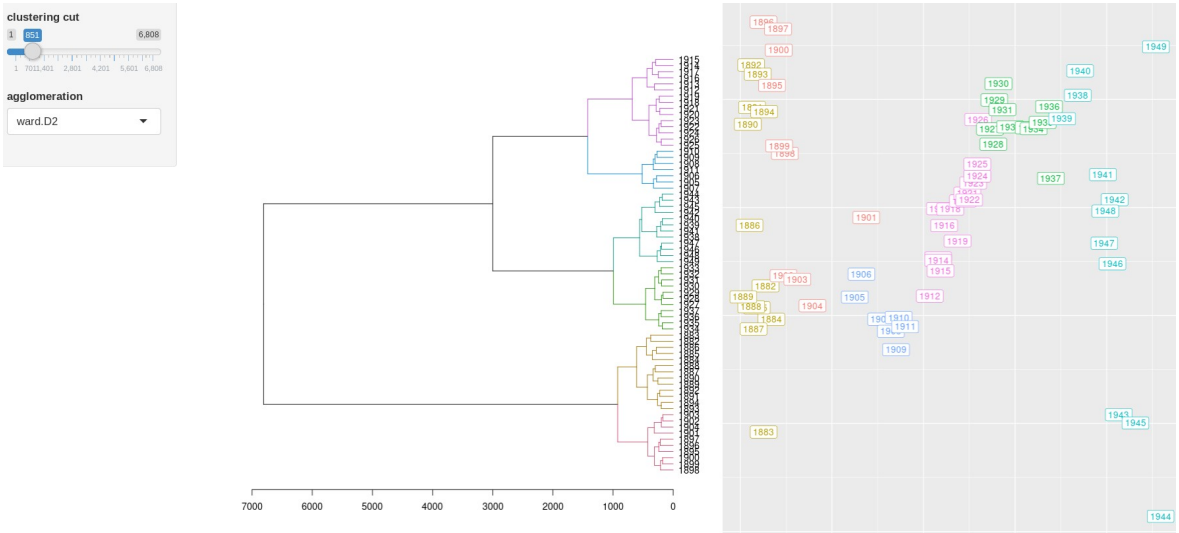


Figure 7: Clustering of years raw of texts. With T-SNE projection on the right. Note : the M-shape is probably the MDS equivalent of the Gluttman effect on FCA (Salem, 1991), typical of ordered sequences. The dendrogram on the other hand shows clearly separated clusters (screenshot from the online interface)

As we can see, this method yield quite sound results, consistent with the historical outline given in Section 2. considering the abolition of imperial examination in 1904, the end of the Qing dynasty in 1911 and the actual institutionalization of the national language in the late 20's. I can now use this clustering results to define more **homogeneous** subcorpora and apply distributional analysis to describe properties of the language(s) in use. Considering the small conflicts in the clustering, I discard small parts of the corpus which may be "too fuzzy". I will start the analysis in 1895 and skip the 1922~1926 data.

7. DISTRIBUTIONAL ANALYSIS OF SOME LEXICAL ITEMS

In this section, I propose a novel methodology for which I developed a set of tools to track the evolution of sets of competing lexical items ("words") through the corpus. To be able to observe contrasts, I select items that dictionaries may point as synonyms or closely related but that the literature points as characteristic of different variants of mandarin, other regional languages or more classical Chinese. I draw these sets mostly from (Coblin, 2000) and complete them by consulting dictionaries.

Traditional methods in corpus linguistics or textometry are not straightforward to apply on our corpus. It consists in a collection of timestamped issues, but with no word segmentation, an imperfect article splitting and not annotation. In this part of the work I decided to experiment a different strategy based on very recent advances in NLP. I rely on contextualized embeddings to perform a distributional analysis of the different usages of the various items. (See 7.1)

Concerning the choice of competing items, I turned to grammatical lexemes (or 虛詞, 'empty words'). Such words are more likely to show the distinction among related languages, as they are less often borrowings or cognates. Consider the important part of shared lexicon between Modern Standard Chinese, other sinitic languages or even Korean and Japanese ; grammatical markers are more often characteristic of one language. (The same goes for example among romance languages). I define the different sets based on suggestions found in (Coblin, 2000)

7.1 Words and Vectors

Distributional analysis is on of the main tools of linguistic studies. It has known multiple computational formulations, mostly resulting in estimating a mapping of words into a vector space, which used to be grouped under the term "distributional semantics", and to which "word embeddings" can be seen as simply the more recent (and computationally efficient) formulation based on neural networks. The main idea is always to build a representation of words as continuous vectors of numbers. This change of representation from discrete strings of characters to vectors allows for all kinds of algebraic computation and facilitate machine learning in many ways (notably by reducing the issue of sparsity). Those methods however typically yield a single vector for each word-type, by averaging all its occurrences. This is problematic as it fail to account properly for polysemy and other kinds of ambiguities, and is dramatic if one wants to see the evolution of a word. One solution, following (Hamilton et al., 2016) could be to further split the subcorpus into different periods and compare the embeddings of a word between subcorpora. The issue is then to align vector spaces of different periods of time. Another issue of further splitting the corpus is also the size of the final split used to train the embeddings, which may become too small in some cases.

The solution I adopted for this study is based on a recent type of neural-based embedding proposed in (akbik et al., 2018), which is designed to take the context into account, called "Contextual String Embeddings" (CSE). With CSE, we can obtain a specific vector for each occurrence of any word. Our own work (Blouin & Magistry, to appear) in which we applied CSE to the task of Name Entity Recognition in Modern Chinese has shown CSE being able to provide good vector representations for Chinese script with or without word segmentation. And it appears to be very well suited for the present study.

The vector representation from CSE is obtained by taking the hidden states of two LSTM language models, one reading forward from the beginning of the left context to the end of the targeted item, and another reading backward from the end of the right context to the beginning of the targeted item. In this way, the whole context has been taken into account and the representation is focused on a specific occurrence of a specific word, the word itself can be of any length, it has not to be a single sinogram. Technical details can be

found in the aforementioned papers, what is specific to this study is that I work on non-segmented Chinese script, without either word nor sentence boundaries. To face this situation, I train sinogram-based LSTMs that simply reads characters, and I keep a context of a fixed length (8).

For training, I randomly sample n-grams sequences of sinograms from each period. To build the item sets to be studied, a sample (at most) 2000 occurrences of each item in various contexts from the same period, and compute the corresponding vectors. The result is a vector space which contains thousands of points, one for each sampled occurrence of each item, all in the same space whithout the need of of alignment.

It is then possible to apply hierarchical clustering and dimension reduction (using T-SNE) to have a big picture of each item sets at each period. I also keep the link to the original text in order to be able to know from which sample comes any data point. The result of this analysis method is provided with an interactive interface on the companion website. By selecting a set of competing items and a period, one can visualize the clustering and the vector space in two dimensions after applying a T-SNE dimension reduction. Coloration of points can be made accordingly to the lexical items or to the clustering (number of cluster is also a user parameter by selecting the height of the cut in the dendrogram on the left). It is also possible to select points and see the text snippets as a concordance under the graphic or to see the text of a specific point with a simple mouseover.

A slider was added on top of interface to filter the points by the year of occurrence. However, this functionality seems to be useful only to confirm that the time-based clustering from the previous section was correct.

7.2 很/狠/甚/什/極

The first set of items focuses on intensifiers for stative verbs. The more frequent one in MSC is 很 hěn. In our corpus it compete with the graphical variant 狠. Matthews in his 1937 dictionary consider the two interchangeable, which is no more the case in MSC. Coblin (2000) points that 很/狠 is typical from Northern Mandarin, including Pekinese. It contrasts with 甚 shèn and 極 jí, which are found in southern Mandarin. Coblin states that for these items the shift from South to North in the Qing koiné is as late as the 19th century. We will see that in the Shun-pao, this shift occurs even later. To complete the picture, I added to the computation the form 什, which is recorded as a graphical variant for 甚.

I invite reader to explore the results through the online interactive figures. The procedure I followed is to first look at the big picture provided by the T-SNE panel, dense regions will contain similar uses of certain items. It is especially efficient to spot frequent multi-sinogram expressions. Selecting zone allows one to retrieve the text for closer reading. I then check the clustering panel with the dendrogram to know the order of grouping and the degree of similarity between groups (the dimension reduction applied to produce the graphic lost a part of this information). The web interface is illustrated on Figure 8 with the clustering of the first period of this set.

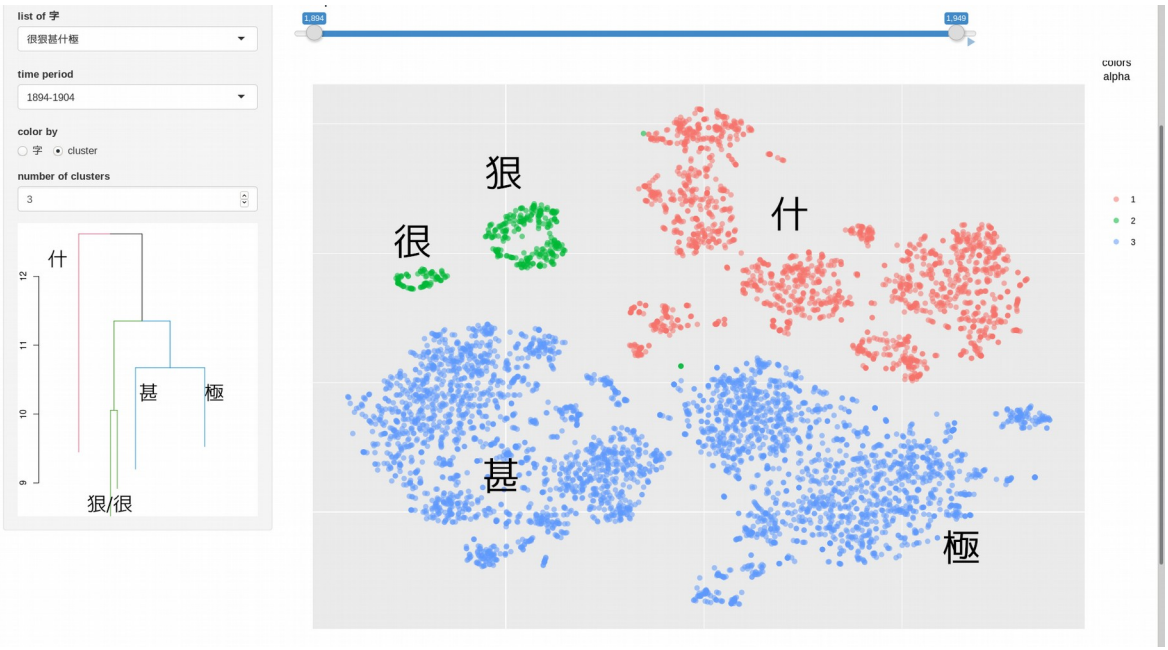


Figure 8: Interface for Context String Embeddings clustering, on the first period with the first set of items

We can see that in the first two periods (1894~1904 and 1905~1911) 狠/很 are rare, with 狠 being more frequent. They seems interchangeable and are rapidly clustered together. In the third period (1912~1922) they become much more frequent and continue to form a consistent cluster. In the fourth period (1927~1938), 很 becomes by far the most frequent form. 狠 starts to form its own cluster, its adverbial usage can still be observed but it tends to be specialized in its 'fierce, cruel' meaning as a stative verb or in compound words, (especially when duplicated 狠狠). This tendency is confirmed in the last period (1939~1949).

We can also see that 什 form a cluster of its own over all the different periods. Beside a small number of occurrences of 甚麼 (starting only in 1927), 什 is almost never a variant for 甚. But its presence in the set lead to other interesting observations. 什 is typically not used in autonomy, but rather in compound forms, even in the early periods of the corpus. For the first period, these expressions include 什物 (goods, things), 什長 (a title in the Qing bureaucracy), 什一之利 ("10% profits", where 什 is actually the variant for 十 ten), 喀什噶爾 ('Kachgar'), 戈什哈 (a Manchu term to design an official guard), in these cases 什 is actually pronounced *shí*.

Interestingly, although the model is character-based, with no a priori knowledge of word boundaries, the CSE+ T-SNE visualization is able to group together the occurrences of such complex expressions. The form 什麼 (or with the variant 麼, 'what') appears in 1905 as a small but distinct group with T-SNE, and it becomes the main use of 什 in the fourth and fifth periods (after 1927).

Concerning 極 and 甚 as intensifier, we can observe that if their frequencies remain high on the five periods, their usages evolve from autonomous forms which combine rather freely to a multitude of compound expressions.

7.2 無 / 未曾 / 不曾 / 沒有

Another case suggested in (Coblin, 2000) is the perfective negative and existential negative. For this function, previous reports on which Coblin based his work claim that 未曾, 不曾 and 沒有 are equivalent and that Mandarin allows for an "apparent freedom of choice". Coblin stresses that the first two are now associated with central dialects of Mandarin while the third one is typically northern. This last one is also

given as an equivalent of the more classical 無 wu by many dictionaries, which also give 未嘗 as a near-synonym of 未曾. As I noticed a confusion between 沒有 and the simplified form 没有 (which came from the transcription and not the original documents), I decided to use both in the set of items under study. We can observe that the algorithms successfully considers these two forms as similar.

The picture of the Shun-pao we can obtain with the proposed algorithms is mostly consistent with Coblin's description but slightly more complex. In the first period (1894~1904), 沒有 and 不曾 are almost nonexistent and we can see two clear cluster, one for 無 and one for the other less classical forms. In the second period (1905~1911) 沒有 is becoming more frequent and constitute a cluster on its own, distant from the other forms that quickly split into the two same clusters as in the previous period. In the 1912~1922 period, we can observe three very distinct clusters: one for 沒有, one for 無 and one for the other forms (with slightly more 不曾 than before). In the last two periods starting in 1927, we are back to a two cluster situation with 無 on its own and 沒有 merging with the other alternatives. The sequence of corresponding dendrogram is presented on Figure 9.

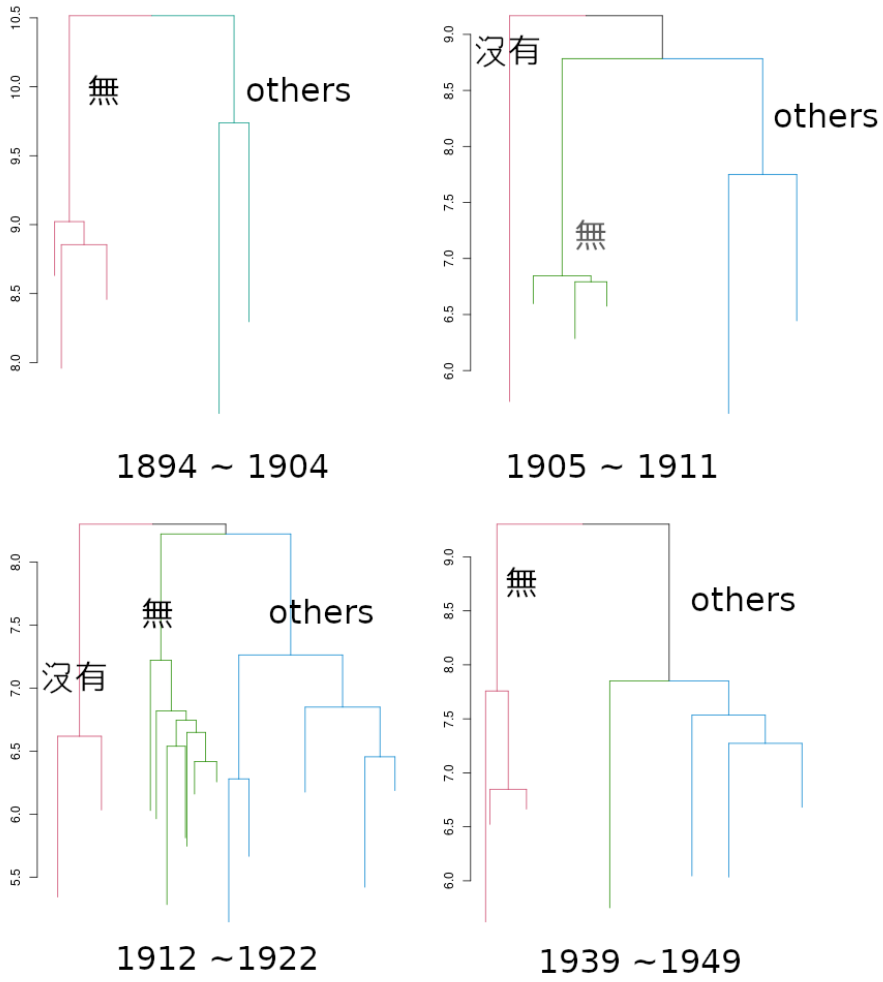


Figure 9: Clustering at different periods to illustrate the evolution of 沒有

This seems to tell a story of a Northern variant appearing later, first as a foreign form, which later merge with the other forms of middle and southern vernaculars. At the end 沒有 is by far the most frequent option and 不曾, which is overall not very frequent seems to have rise in parallel with 沒有.

7.3 Comments on the proposed method

More case-studies will be needed to obtain the full description of our object, but the two sets of items proposed here are sufficient to see the advantages and limitations of the method. Relying on Context String Embeddings, hierarchical clustering and T-SNE allows me to overcome the main issues of other methods including n-grams counts and word embeddings. The first one is to tackle the ambiguity which can arise from homonymy, polysemy or evolution in usages. This issue is worsened by the unsegmented nature of the Chinese script, especially in the case of single sinograms that can occur in very different, more or less frozen expressions. Consider for exemple the sinogram 無, observing its frequency of use or a single vector for all of its occurrences would not tell much about its diversity of usages and its evolution. The same goes for the progressive specialization of 狠. The second issue this method does not suffer from is the need for spaces alignment. As it build a single space with all the occurrences, we don't need to find a way to align the different years in each period.

On the other hand, and just like more classical word embeddings, this method involves a lots of parameters such as the number of dimensions, the window size and the training algorithm. In this experiment I simply adopted some default parameters which have shown good result in our previous work on NER in Chinese. These defaults already provided useful output, but it would deserve a more complete exploration. We can see that the CSE are still highly sensitive to the targeted wordform. It was able to group together interchangable variants in some contexts like in 甚巨(/鉅) or "無綫(/線)電臺", but it still distinguishes clearly between apparently similar usages of 狠 and 很. It is very likely that a different set of parameters could give even more satisfying results by balancing the relative weights of the context and the targeted form.

8. CONCLUSION

In this paper, I argued that the Shun-pao as a corpus offers a unique window on language(s) in use at a very special time for MSC history. I proposed computational methods and tools to face its large size without the need of prior hypothesis regarding its inner structure, unraveling fluctuation in punctuation practices, and a tentative data-driven periodization. I tried to demonstrate that CSE combined with clustering and dimension reduction provide a great tool to explore a corpus and to go more straightforwardly to interesting phenomenon despite the huge size of the corpus under scrutiny. All of this was done by providing an interactive user interface so the results can be checked and further explored by the reader.

On top of the periodization, the main finding is to confirm the heterogeneity and complexity of the corpus. But also to show that this complexity can be dealt with confidently using CSE. It seems that consistently with Simmons (2017)'s claim, the actual linguistic practices observable in our corpus remain all but standardized until its end in 1949. One of the main limitation of this study is the granularity, which is year or "period" based. This is due to the provided meta-data which is reliable only down to the "issue" (daily) level. I jump too quickly from a very distant reading at the period level to a (too) close reading of short snippets of text. What we can observe in this way underlines the need for a proper article segmentation. Different and competing usages can be observed on the same day (in terms of punctuation or language) but I can not say if they coexist in one same article. This level is likely to provide an important and consistent level of analysis. It will also be relevant for the sake of information retrieval.

The next natural step will be to work on such article segmentation, which could be addressed as a joint task of segmentation and clustering of the articles. Such work will benefit from our experience with the CSE which can provide an efficient first step of unsupervised pre-training.

ACKNOWLEDGEMENT (致謝)

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 788476)

REFERENCES (參考文獻)

Akbik, A., Blythe, D., & Vollgraf, R. (2018). Contextual String Embeddings for Sequence Labeling. In Proceedings of the 27th International Conference on Computational Linguistics (pp. 1638–1649). Santa Fe, New Mexico, USA: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/C18-1139

Bamman, D., Popat, S., & Shen, S. (2019). An annotated dataset of literary entities. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 2138–2144). Minneapolis, Minnesota: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1220

Chen, M. (1970). The Chinese language today; features of an emerging standard: Paul Kratochvíl Modern Languages and Literature. Hutchinson University Library, London 1968. 199 pp. 15s (cased 35s). Lingua, 25, 82–89. https://doi.org/10.1016/0024-3841(70)90022-7

Coblin, W. S. (2000). A Brief History of Mandarin. Journal of the American Oriental Society, 120(4), 537–552. https://doi.org/10.2307/606615

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1489–1501). Berlin, Germany: Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1141

Heafield, K., Pouzyrevsky, I., Clark, J. H., & Koehn, P. (2013). Scalable Modified Kneser-Ney Language Model Estimation. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 690–696). Sofia, Bulgaria: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P13-2121

Huang, C.-R., Lin, J., Jiang, M., & Xu, H. (2014). Corpus-based Study and Identification of Mandarin Chinese Light Verb Variations. In Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (pp. 1–10). Dublin, Ireland: Association for Computational Linguistics and Dublin City University. https://doi.org/10.3115/v1/W14-5301

Kaske, E. (2008). The Politics of Language in Chinese Education: 1895 - 1919. BRILL.

Press, T. M. (2012, October 28). 1965: Aspects of the Theory of Syntax [Blog]. Retrieved November 4, 2019, from https://mitpress.mit.edu/blog/1965-aspects-theory-syntax

Quote Unquote Language Reform: New-Style Punctuation and the Horizontalization of Chinese. (2017, November 18). Retrieved November 4, 2019, from http://u.osu.edu/mclc/journal/abstracts/mullaney/

Simmons, R. V. (2017). Whence Came Mandarin? Qīng Guānhuà, the Běijīng Dialect, and the National Language Standard in Early Republican China. Journal of the American Oriental Society, 137(1), 63–88. https://doi.org/10.7817/jameroriesoci.137.1.0063

Weng, J. (2018). What Is Mandarin? The Social Project of Language Standardization in Early Republican China. The Journal of Asian Studies, 77(3), 611–633. https://doi.org/10.1017/S0021911818000487

Zampieri, M., Malmasi, S., Scherrer, Y., Samardžić, T., Tyers, F., Silfverberg, M., … Jauhiainen, T. (2019). A Report on the Third VarDial Evaluation Campaign. In Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects (pp. 1–16). TOBEFILLED-Ann Arbor, Michigan: Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-1401