Le(s)? chinois du Shun-pao

Pierre Magistry

Le(s)? chinois du Shun-pao 申報

Pierre Magistry
ertim, INALCO, Paris
pierre.magistry@inalco.fr

MOTS-CLÉS : Corpus historique, AFC, plongements contextualisés, chinois mandarin, chinois classique.

KEYWORDS: Historical corpora, FCA, contextualized word embeddings, Mandarin Chinese, Classical Chinese.

1 Introduction

Les travaux présentés ici poursuivent un objectif double. Premièrement, il s'agit d'approfondir l'étude inédite que constitue l'exploration du premier journal quotidien imprimé en sinogrammes, le Shun-pao, dont le texte intégral numérisé n'était pas disponible pour des traitements massifs jusqu'à tout récemment. Ce corpus est un témoignage d'une pratique du chinois écrit au tournant du 20ème siècle. La période que couvre ce quotidien présente un intérêt tout particulier. C'est à cette époque que se construit et s'impose le projet d'une langue nationale unique pour la Chine, et que les formes vernaculaires prennent le dessus sur l'usage du chinois classique dans les pratiques écrites. L'accès au corpus de la totalité des textes numérisé rend possible des études quantitatives pour caractériser la diversité linguistique du corpus et son évolution sur une période allant de 1872 à 1949 en observant directement les usages.

Le second objectif de notre travail est méthodologique, en proposant une comparaison de plusieurs approches. On souhaite mettre en regard les apports de méthodes neuronales basées sur des plongements lexicaux contextualisés avec une approche plus classique fondée sur l'analyse factorielle des correspondances. On insistera sur les critères pris en compte pour favoriser une méthode plutôt qu'une autre, et sur les types de questionnements que l'on peut aborder avec l'une et l'autre des méthodes.

2 Les corpus

Cette étude se focalise sur le corpus du Shun-pao, mais nous avons aussi recourt à d'autres corpus comme points de comparaison. Il s'agit d'une part d'un corpus arboré de chinois moderne standard de Taïwan tel que pratiqué dans les années 1990 (le Sinica Treebank), et d'autre part d'échantillons d'un corpus historique divisé en trois périodes distinctes (chinois archaïque 上古, chinois médiéval 中古 et chinois pré-moderne 近世).

Le corpus du Shun-pao est un corpus inédit, dont l'accès au texte intégral n'a pu être rendu possible que tout récemment grâce au projet ERC *Elites, Networks and Power in Modern China*. Il contient le

texte de tous les articles (en excluant les publicités) qui ont été OCRisés et corrigés manuellement. Il comporte environ 750 millions de sinogrammes. Une version des données avec découpage en pages et en articles a pu être obtenu dans un second temps, ce qui a fait disparaître certaines limites des travaux précédents. Mais le découpage en articles nous semble encore très imparfait. La typographie est aussi en grande partie perdue, ce qui a posé problème pour l'étude de l'évolution des pratiques de ponctuation, mais n'est pas limitant pour la présente étude.

3 L'approche par « modèles de langue »

Des travaux précédents (Magistry, 2019) à la suite desquels nous nous situons reposent sur l'entraînement de modèles langues à partir du texte brut des documents. Notons que cette stratégie était en partie imposée par l'absence de métadonnées, par une segmentation limité du corpus et par l'absence de points de comparaison avec les autres corpus, invoqués dans un second temps.

Avec une segmentation par année du corpus, il est tout de même possible de proposer une périodisation basée sur la perplexité d'une multitude de modèles de langue (un par année) à la lecture de chaque autre année du corpus. On peut ainsi construire une matrice de distances et appliquer du clustering hiérarchique pour repérer des points de rupture.

De plus, en utilisant des plongements lexicaux contextualisés obtenus par l'entraînement de modèles de langues avec des BiLSTM (Akbik et al., 2018), on a pu suivre l'évolution de l'usage de certains items lexicaux choisis dans la littérature de linguistique historique (Peyraube, 1996; Coblin, 2000)

4 L'analyse des correspondances

Après l'obtention de corpus comparables et d'une segmentation du corpus plus fine (a minima à la page), on peut recourir à l'analyse factorielle des correspondances (Benzécri, 1973). On dispose ainsi d'une méthode éprouvée pour situer la langue et visualiser son évolution en se basant sur des comptes d'occurrences de mots « vides » (ou grammaticaux), que l'on estime plus caractéristiques de la langue d'un texte que les mots « pleins » qui peuvent plus facilement être des emprunts ou être surtout caractéristiques des thèmes et du genre d'un texte. On peut ainsi obtenir des plans factoriels qui situent des pages de notre corpus par rapport aux chinois anciens et modernes.

5 Points étudiés

On étudie en particulier certains facteurs et propriétés des modèles pour dégager leurs atouts respectifs. Il s'agit des propriétés des corpus utilisés en taille, en qualité des annotations, en finesse des métadonnées et des objets mis au centre des analyses (ensemble de textes, document unique, type ou occurrence token). On observe aussi des comportements différents face aux variantes graphiques des sinogrammes, qui n'ont été standardisés que récemment et au problème de la segmentation en mots.

On montrera qu'en combinant ces méthodes, il envisageable d'enrichir notre corpus en catégorisant les textes qui le composent et aussi de suivre des évolution diachronique de l'emploi de certains items lexicaux.

Références

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Benzécri, J.-P. (1973). *L'analyse des données 2 l'analyse des correspondances*. #0, Dunod, Paris, Bruxelles, Montréal.

Coblin, W. S. (2000). A brief history of mandarin. *Journal of the American Oriental Society*, 120(4) :537–552.

Magistry, P. (2019). Languages(s) of the SHUN-PAO, a Computational Linguistics account. In *10th International Conference of Digital Archives and Digital Humanities*, Taipei, Taiwan.

Peyraube, A. (1996). Recent issues in chinese historical syntax. In Huang, C.-T. J. and Li, Y.-h. A., editors, *New Horizons in Chinese Linguistics*, pages 161–213. Springer Netherlands, Dordrecht.