Un état des lieux du traitement automatique du mandarin

Laurent Prevot, Pierre Magistry, Chu-Ren Huang

Un état des lieux du traitement automatique du mandarin

Laurent Prévot*& Pierre Magistry & Chu-Ren Huang*****

INTRODUCTION

Après avoir été considérée comme une langue sous-dotée en termes d'outils automatiques et de ressources langagières, le mandarin, tel que standardisé et institutionnalisé comme langue officielle en Chine, à Singapour et à Taïwan, est devenu ces dernières années une des langues concentrant l'essentiel des efforts de la communauté du traitement automatique des langues. Outre la démographie de la communauté académique, deux autres facteurs au moins peuvent expliquer cet engouement: les aspects stratégique[1] liés au traitement automatique du mandarin et, de manière liée mais plus fondamentale, la croissance rapide de la taille des données produites en mandarin. Ce dernier point dans un domaine particulièrement gourmand en données fait du mandarin un nouveau terrain de jeu de premier choix pour les spécialistes du traitement automatique des langues.

Cet état de l'art a été constitué à partir de l'expertise des auteurs concernant en particulier les questions des ressources et des premières étapes de traitement et a été complété par un examen de l'anthologie de l'ACL pour les années 2013 et 2014.

Nous ne traiterons ici que du mandarin standard (普通话/pǔtōnghuà de la République Populaire de Chine, le 新加坡華語/ xīnjiāpō huáyǔ et le 國語/guóyǔ de Taïwan)[2], écrit en caractères chinois (sinogrammes). Le traitement du signal de la parole, comme celui des textes écrits en d'autres langues sinitiques sortent du cadre du présent article.

* Aix Marseille Université, CNRS, LPL UMR 7309, 13100, Aix-en-Provence, France. Courriel : laurent.prevot@univ-amu.fr
** National Taiwan University, Graduate Institute of Linguistics, Taipei, Taiwan & INRIA, Université Paris Diderot, Alpage, Paris, France. Courriel : pmagistry@gmail.com
*** The Hong Kong Polytechnic University, Chinese and Bilingual Studies, Hong Kong, China (SAR). Courriel : churen.huang@polyu.edu.hk

[1] Nous pensons ici à l'importance croissante de cette langue dans les sphères industrielles et économiques.

[2] Il existe des différences entre ces variantes de mandarin (ainsi qu'avec les variantes non standard), mais elles n'affectent pas de façon significative notre propos. Nous utilisons donc le terme simple de «mandarin» dans notre texte pour ne pas alourdir celui-ci.

1. Organisation de l'article

Dans cet article, après avoir introduit le TAL et ses techniques (Section 2), nous allons suivre les différents étages de la chaîne de traitement. Nous commencerons (Section 3) par examiner les problèmes qui se posent premièrement pour la représentation informatique et la saisie de cet objet graphique qu'est le caractère chinois, puis nous passerons au niveau de la morphologie et du lexique (Section 4) avec les tâches de segmentation et d'étiquetage morpho-syntaxique. Les questions des analyses syntaxique puis sémantique et discursive seront abordées dans les sections 5 et 6. Nous aborderons, enfin, dans la section 7 la question cruciale des ressources linguistiques (corpus, lexiques), des outils disponibles librement et de leurs applications.

2. Présentation du TAL

2.1 Le traitement automatique des langues en quelques mots

Le traitement automatique des langues est généralement associé à trois notions pour les non-spécialistes:

le *traitement automatique des langues* (TAL) (pris ici dans un sens étroit correspondant à l'anglais *Natural Language Processing*) qui consiste à la résolution de tâches spécifiques liées au langage naturel via des processus automatiques ;

la *linguistique computationnelle* (*Computational Linguistics*) qui établit des modèles opérationnels pour les questions linguistiques ;[3]

la linguistique outillée (Habert 2004 ; Heiden 2010) qui est essentiellement une linguistique des grands corpus rendue possible par des pré-traitements automatiques et des outils de requête issus du traitement automatique des langues.

Cet article ne constitue pas une introduction exhaustive à ces domaines et nous recommandons aux lecteurs intéressés de consulter les excellents ouvrages d'introduction que sont Jurafsky (2000) pour une présentation généraliste, Manning (1999) pour une focalisation sur les approches statistiques, Bird (2006) pour une présentation plus pratique via l'exemple et, par exemple, Tanguy

[3] TAL et linguistique computationnelle sont généralement tenus pour synonymes y compris parmi les spécialistes. Néanmoins, pour des raisons pédagogiques, nous trouvons utile la distinction opérée ici. Bien que recouvrant en grande partie les mêmes méthodes et les mêmes sujets, nous considérerons que le TAL a une finalité plus appliquée — et souvent associée au terme d'*ingénierie linguistique* (avec des tâches définies comme telles, la traduction automatique ou le résumé de document) — que la linguistique computationnelle qui cherche à résoudre les mêmes questions que les autres paradigmes de la linguistique (comme l'analyse syntaxique ou la résolution anaphorique).

(2007) pour un ouvrage pratique en français sur l'utilisation de techniques avancées de linguistique outillée. Enfin, en ce qui concerne plus spécifiquement le traitement du mandarin, l'ouvrage d'introduction de Wong (2009) propose un point d'entrée dans ce domaine.

Les différents niveaux du traitement automatique sont traditionnellement présentés dans une chaîne de traitement correspondant au niveau de l'analyse linguistique : segmentation en unités (*tokenisation*), étiquetage morphosyntaxique, analyse syntaxique, analyse sémantique, analyse discursive. Cette vision est cependant simplificatrice en raison, notamment, de l'émergence de modèles joints qui visent à la résolution de plusieurs de ces tâches de manière simultanée. Une telle approche jointe permet de mieux prendre en compte les phénomènes de dépendances entre niveaux.

2.2. Différentes manières de faire du TAL

Les méthodes utilisées en TAL peuvent être classées sur un continuum entre des méthodes fondées uniquement sur des connaissances d'experts (linguistes en particulier) et des méthodes d'induction n'utilisant initialement que des données brutes (non analysées), mais en très grande quantité. On distingue cependant deux grandes familles de méthodes : les méthodes dites «symboliques» ou à base de règles qui sont définies par des experts (*rule-based*) et les méthodes dites «statistiques» reposant sur l'apprentissage automatique (*machine learning*).

Les méthodes de la première famille utilisent des programmes qui appliquent de façon systématique et automatique un ensemble de principes définis le plus souvent manuellement par un expert ou un ensemble d'experts. Les règles peuvent être écrites dans différents formalismes. Un paradigme souvent utilisé est celui des transducteurs à états finis qui permettent de réécrire une séquence de symboles en une nouvelle séquence. Par exemple pour passer d'un texte brut à un texte enrichi d'informations linguistiques comme les parties du discours. D'autres systèmes à base de règles proposent des mises en œuvre de grammaires décrites dans un langage issu de formalismes linguistiques (par exemple xLFG pour le formalisme LFG (Lexical Functional Grammar)[4], ou LKB pour la HPSG (Head-driven Phrase Structure Grammar)[5]...). Il faut noter que ces approches tendent à s'hybrider et à intégrer par exemple des informations probabilistes (recueillies sur corpus) afin d'affiner leurs principes.

Les méthodes de traitement par apprentissage automatique permettent d'éviter d'avoir à écrire manuellement les grammaires ou autres règles d'analyse. Au lieu de cela elles utilisent généralement des corpus annotés (par des humains) pour entraîner un programme qui devra ensuite être capable d'imiter les prises de décision des annotateurs du corpus d'entraînement afin de pouvoir annoter de nouvelles données. Il faut donc disposer de corpus annotés manuellement pour

[4] http://www.xlfg.org/
[5] http://moin.delph-in.net/LkbTrollet

entraîner le programme et trouver des algorithmes capables de reproduire de telles annotations à partir des exemples donnés en entraînement.

Certaines méthodes d'apprentissage automatique ne nécessitent cependant pas de données manuellement annotées pour leur entraînement. Elles analysent directement des données «brutes» dans lesquelles elles identifient des régularités statistiques utiles pour la tâche. Ces méthodes sont dites *non supervisées*, par opposition aux méthodes d'apprentissage *supervisées* qui ont recours à l'annotation manuelle. Les méthodes non supervisées étant nettement moins répandues que les méthodes supervisées, lorsque le type de supervision n'est pas précisé nous référons ici à l'apprentissage supervisé.

Les méthodes à base de règles présentent l'avantage de ne pas nécessiter de grosses quantités de données (annotées ou non), et peuvent commencer à être utilisées dès l'écriture de la première règle. Cependant, il s'avère très difficile de rédiger un jeu de règles complet, traitant l'ensemble des phénomènes linguistiques d'une langue. Ces méthodes ont donc un problème de couverture. Les campagnes d'évaluation rendues possibles par la diffusion de corpus annotés classent le plus souvent les méthodes à base de règles derrière les méthodes statistiques. De nos jours, elles restent utilisées pour mettre à l'épreuve des théories de linguistique formelle et aider à les développer en validant leur cohérence. On les trouve aussi très utilisées pour certaines tâches très spécifiques comme les pré-traitements (voir par exemple Unitex[6], ou SxPipe, Sagot 2008).

Les méthodes d'apprentissage non supervisées se révèlent, elles aussi, moins efficaces que les méthodes supervisées, car elles ne bénéficient pas des indications fournies par les annotations des corpus. Par ailleurs, il est parfois difficile de faire correspondre le résultat de leurs analyses avec les annotations issues de telle ou telle théorie linguistique, ce qui rend délicates leur évaluation et leur utilisation dans une chaîne de traitement. Elles peuvent toutefois être vues comme un moyen d'observer les données pour établir de nouvelles théories. Elles sont aussi utilisées pour fournir des indications aidant l'apprentissage des méthodes supervisées.

Les hybridations évoquées entre les deux grandes approches viennent de multiples propositions.

Premièrement, l'apprentissage supervisé repose sur une forme d'analyse humaine (plus ou moins experte selon la nature des annotations exploitées). Il faut donc considérer cette technique comme hybride, car elle n'est pas uniquement basée sur l'induction à partir de données brutes; des principes, voire des règles ont été fournis aux annotateurs / analystes. Dans ce cas, on peut considérer que l'ensemble des règles utilisées est limité aux règles suffisamment robustes pour être appliquées systématiquement par un groupe d'annotateurs

[6] http://www-igm.univ-mlv.fr/~unitex/

tandis que le moteur d'apprentissage cherche les meilleures combinaisons de ces règles pour trouver la solution du problème.

Deuxièmement, les principaux algorithmes d'apprentissage statistiques utilisés à ce jour (par exemple les SVM -- Support Verctor Machine -- et les CRF --Conditional Random Fields--) permettent de spécifier des noyaux informatifs qui sont utilisés par le moteur d'apprentissage. Tellier (2009) appelle ces noyaux des atomes de connaissances et ces derniers proviennent des connaissances des experts. Dans ce dernier cas, les connaissances expertes (linguistiques) sont des contraintes générales dans lesquelles les moteurs statistiques vont se couler pour inférer toutes les subtilités présentes dans les données et pour lesquelles il est peu réaliste d'imaginer un ensemble de règles prédéfinies. (Voir Tellier 2009 et Boitet 2008 pour une discussion plus détaillée de ces éléments).

En présence d'un grand jeu de données[7], il est difficile de contester la suprématie des méthodes d'apprentissage supervisé, mais les autres méthodes ne peuvent être ignorées. Par ailleurs, la nécessaire disponibilité d'un corpus annoté pour l'entraînement qui soit de taille suffisante et de nature correspondant aux données à traiter (même genre, même dialecte, même médium) peut se révéler problématique. Dans ce cas les méthodes à base de connaissances expertes sont les plus performantes comme illustré dans Raghunathan (2010). La question de l'adaptation à un nouveau domaine (*domain adaptation*) d'un modèle déjà entraîné sur des données différant plus ou moins grandement des données à traiter est une des questions fréquemment abordées dans les travaux récents en TAL, toutes tâches confondues.

2.3 Les spécificités du mandarin

Étant donné que les caractéristiques du mandarin écrit et parlé diffèrent radicalement des langues européennes, le traitement automatique du mandarin a du faire face à de nouveaux défis. Par exemple, les deux premières difficultés rencontrées dès le début du traitement automatique sont l'absence d'espaces entre les mots et la faiblesse des marques de morphologie explicite.

Avec son système d'écriture fondé sur le sinogramme, le mandarin ne marque pas explicitement les frontières entre mots typographiques. La segmentation en mots (ou *tokens*), brique de base de la quasi-totalité des applications de traitement automatique est considérée comme une phase de pré-traitement relativement triviale pour des langues comme l'anglais ou le français. Elle

[7] Il est difficile de dire en toute généralité ce qu'est un grand jeu de données mais l'immense majorité des travaux dans ce domaine travaille avec des nombres d'exemples (points de décision) allant de plusieurs dizaines de milliers à plusieurs millions. L'apprentissage non supervisé nécessitant généralement des jeux de données bien plus grands encore.

présente en revanche un réel défi pour le traitement des textes écrits en mandarin. Deuxièmement, le faible nombre de marques morphologiques et la possibilité souvent offerte d'effectuer des dérivations non marquées rendent l'analyse syntaxique plus malaisée qu'à l'accoutumée.

3. DU PINCEAU AU CLAVIER

L'écriture chinoise (entendue comme système graphique basé sur les caractères chinois, pouvant transcrire différentes langues) possède des spécificités qui représentent autant de défis pour sa numérisation et son traitement automatique.

Même si en se focalisant ici sur son usage pour la transcription du mandarin (ou «chinois moderne standard»), l'écriture chinoise est le résultat d'une longue histoire qui a vu différents usages rivaliser au fil de l'histoire ou d'une région à une autre. Retracer cette histoire sort largement du cadre de cet article, mais nous pouvons tout de même noter les particularités, issues de ces divergences, qui peuvent affecter le traitement automatique du mandarin standard actuel.

Si l'unité graphique la plus saillante est le caractère, il est généralement possible de subdiviser un caractère en éléments composants. Par exemple nous pourrons reconnaître dans le caractère 語 les éléments 言 et 吾, le second étant lui-même composé de 五 et de 口. Cette possible décomposition est ignorée des standards d'encodage actuels qui affectent un code unique à un caractère entier (語 est codé U+8A9E en Unicode), mais elle est à la base de la formation de l'écrasante majorité des caractères, en jouant sur une homophonie entre le caractère créé et un de ses composants, Sagart (2006). L'écriture manuscrite n'étant pas contrainte par les questions d'encodage, différentes variantes d'un même caractère[8] ont pu être utilisées en différents lieux ou époques. Malgré les efforts de standardisation, statuer sur la question des variantes peut parfois être délicat et cela se traduit par la coexistence de caractères proches jusque dans certains encodages informatiques. Par exemple le caractère 峰 (fēng) désignant le sommet (d'une montagne) est une variante courante de 峯, les deux sont la plupart du temps interchangeables (sauf dans le cas des noms propres). Nous remarquons qu'ils ont les mêmes éléments composants, mais que ceux-ci sont disposés de manières différentes. Pour cet exemple, l'Unicode utilise deux codes différents. Ceci peut conduire à des problèmes de couverture des ressources, un lexique pourrait inclure l'entrée 峯會 'fēnghuì' *rencontre au sommet* et ignorer la variante graphique 峰會. Le standard Unicode inclut des informations concernant les variantes graphiques, mais celles-ci sont toujours sujettes à débat. Cette question est d'autant plus problématique que l'Unicode vise à décrire les caractères chinois indépendamment de la langue qu'ils transcrivent. Se pose

[8] Parler de variantes d'un caractère est déjà une analyse centrée sur le caractère. Au départ il s'agit de plusieurs façons similaires de transcrire le même morphème ou sa traduction.

alors la question des variantes d'une langue à l'autre, qui est théoriquement difficile et peut heurter les sensibilités nationales.

La décennie précédente a également vu la mise à disposition de ressources proposant une description de la décomposition des caractères, par exemple par le projet CHISE[9] qui utilise le système IDS préconisé par le consortium Unicode (Allen 2014), le 小學堂/ *xiǎoxuétáng* de l'Academia Sinica[10], qui intègre les données du projet[11] ou le CDL (*Character Description Language*) proposé par l'Institut Wenlin (Bishop 2003). Elles permettent d'ouvrir de nouveaux champs d'études et d'aider à la résolution de nouvelles tâches comme la correction orthographique automatique en mandarin (Yu et al. 2014).

4. MORPHOLOGIE ET LEXIQUE

4.1 Segmentation

Comme évoqué plus haut, la segmentation en mots est une première étape essentielle pour le traitement automatique. Cette tâche n'est pas triviale en mandarin en raison notamment du vague entourant la notion de mot dans cette langue. Différents guides de segmentation pour le mandarin ont été proposés par les équipes qui ont constitué les principaux corpus segmentés manuellement. Aucun de ces guides ne s'est cependant imposé comme standard. Au contraire, il est généralement admis, Wu (2003), que des objectifs applicatifs différents peuvent nécessiter des segmentations différentes. Le groupe de travail ACL-SIGHAN (*Association for Computational Linguistics - Special Interest Group in Chinese Language Processing*) a organisé les premières compétitions internationales en proposant plusieurs corpus de référence afin de permettre une évaluation objective. Ces corpus ont été construits à partir de jeux de données employant différents schémas d'annotation et donc de segmentation (Academia Sinica, Taiwan ; City University of Hong Kong, Peking University, University of Pennsylvania, Microsoft Research).

La segmentation en mots implique deux questions de recherche principales: la désambiguïsation des frontières de mots et la reconnaissance des nouveaux mots (noms propres, néologismes). Ces deux questions peuvent être travaillées de manière indépendante ou simultanée. Les techniques de segmentation du mandarin peuvent être classées dans trois grandes familles: (i) fondées sur les dictionnaires, (ii) statistiques, (iii) hybrides. Les performances des premières dépendent grandement de la couverture des dictionnaires en question, couverture qui de toute manière laisse toujours à désirer en raison des multiples sources

[9] http://www.chise.org/
[10] http://xiaoxue.iis.sinica.edu.tw/
[11] http://cdp.sinica.edu.tw/

d'items hors-vocabulaire (*OOV -- out-of-vocabulary*). Par conséquent, la plupart des systèmes à base de dictionnaire incluent un module d'identification des mots inconnus et une heuristique de résolution des cas ambigus.

L'émergence de grands corpus et l'avènement de l'apprentissage automatique ont permis le développement de techniques statistiques pour la segmentation automatique. Xue et al. (2003) ont proposé une méthode supervisée très efficace fondée sur les séquences de caractères. La méthode consiste en la génération d'une étiquette (indiquant que le caractère se trouve au début, à la fin ou à l'intérieur d'une unité de segmentation) pour chaque caractère d'un texte. La séquence de ces étiquettes est ensuite examinée afin d'optimiser les placements de frontières. Cette méthode a ensuite été déclinée pour une grande variété de systèmes d'apprentissage automatique (en particulier les *Conditional Random Fields* et les *Support Vector Machines*).

4.2 Évaluation

L'évaluation des systèmes de segmentation se fait en comparant la sortie des systèmes automatique à une annotation manuelle de référence. Inspirée par le domaine de la *recherche d'information*, la qualité est estimée par des mesures de *précision*, *rappel* et *f-mesure* sur deux types d'objets: les coupures et les mots. La précision est le ratio du nombre de coupures (resp. de mots) correctement identifiées par les systèmes sur les nombres de coupures (resp. de mots) reconnus. Le rappel est le ratio du nombre de coupures (resp. de mots) correctement identifiées par les systèmes sur les nombres de coupures (resp. de mots) présents dans l'annotation de référence. La f-mesure est la moyenne harmonique des deux mesures précédentes. Les scores sur les coupures sont plus faciles à interpréter en eux-mêmes, mais les scores sur les mots sont susceptibles d'être une meilleure indication du taux d'erreur qui va se propager vers les niveaux d'analyse plus profonds. L'état de l'art actuel se situe aux alentours de 97 % de f-mesure sur les mots avec, par exemple, le système de ICTCLAS de Zhang (2003), qui adopte une méthodologie à base de modèles de Markov en cascade.

Il est cependant important de souligner que les meilleurs résultats sont obtenus dans des conditions idéales pour de l'apprentissage automatique supervisé où les données d'entraînement sont disponibles en grande quantité et sont de nature très proche des données utilisées pour l'évaluation. Dans bien des cas pratiques, de tels jeux d'entraînements ne sont pas disponibles et l'on ne peut espérer d'aussi bons résultats.

Une autre limite notable est que beaucoup d'efforts sont consacrés à améliorer les systèmes de segmentation, mais les questions linguistiques sous-jacentes à l'annotation des corpus utilisés pour l'entraînement et l'évaluation restent peu explorées. La segmentation du mandarin en «mots» vise à ramener la problématique originale posée par l'écriture chinoise au cas mieux connu en

TAL des écritures qui marquent typographiquement des frontières de mots standardisées et connues des scripteurs par une normalisation de l'écriture (comme c'est le cas pour le français ou l'anglais). Cependant, les normes de segmentation proposées par les différentes équipes ayant produit les corpus sont peu comparables à la standardisation orthographique du français qui est imposée politiquement à l'ensemble des locuteurs lettrés de la langue. Les guides de segmentation des corpus en mandarin ne sont suivis que par les annotateurs et n'ont aucune influence sur les pratiques des locuteurs.

Parallèlement aux travaux sur la segmentation du chinois, les spécialistes du TAL s'intéressent de plus en plus aux cas où la standardisation typographique ne correspond pas à l'unité la plus pertinente pour l'analyse linguistique (le cas des expressions dites polylexicales) et aux cas où la pratique écrite ne suit pas les normes typographiques («écrit spontané» comme dans le cas des SMS). Malgré les fortes similitudes de ces problématiques avec celles de la segmentation, nous ne relevons aujourd'hui que trop peu de contacts entre les différentes communautés de chercheurs concernées (Magistry 2013).

4.3 Étiquetage morpho-syntaxique

Les ambiguïtés d'étiquetage morpho-syntaxique sont particulièrement sévères pour le mandarin en raison de la faiblesse de sa morphologie explicite, information cruciale dans l'identification de la fonction syntaxique. Les systèmes de l'état de l'art sont à 95 % de précision. Globalement les techniques d'étiquetage pour le mandarin ne diffèrent pas profondément de celles pour les autres langues. Cependant, les erreurs provenant de la segmentation se propagent sur l'étiquetage. Les méthodes jointes (réalisant simultanément la segmentation et l'étiquetage) ont été développées principalement avec des méthodes d'apprentissage automatique supervisées. Elles sont plus performantes que celles où les deux tâches sont réalisées en série, car elles évitent la propagation des erreurs et parviennent à exploiter des prédictions d'étiquettes pour aider à la segmentation. Lorsque le texte en entrée de l'étiqueteur n'est pas préalablement parfaitement segmenté, la précision de l'étiquetage chute d'environ 5 points. Les résultats rapportés dans les récentes études varient aussi grandement d'un corpus à l'autre.

4.4 Désambiguïsation sémantique

La désambiguïsation sémantique (*word sense disambiguation*) consiste à associer à chaque mot en contexte une des acceptions associées à son entrée lexicale dans un dictionnaire ou un thésaurus. Le choix est trivial en cas de mot univoque, mais peut se révéler très complexe dans les cas fréquents de polysémie. C'est une tâche cruciale pour qui veut rentrer dans le sens du texte. Elle a par conséquent reçu une attention considérable et à fait l'objet de multiples

campagnes d'évaluation (Kilgarriff & Rosenzweig 2000). Concernant le mandarin, Dang et al. (2002) montrent que pour la désambiguïsation des verbes[12] la prise en compte d'informations syntaxiques de niveau intermédiaire (comme la constituance) n'est pas aussi utile que pour l'anglais. L'hypothèse évoquée est que les arguments sont plus systématiquement juxtaposés aux verbes qu'en anglais. L'information de collocation suffit donc généralement, à condition toutefois que la segmentation ait bien été effectuée. En effet, ces auteurs soulignent que bien des désambiguïsations sont nécessaires au moment de la segmentation, elles ne sont donc plus à effectuer plus tard.

5. ANALYSE SYNTAXIQUE

Une fois l'analyse morpho-syntaxique effectuée, la tâche d'analyse syntaxique est formellement assez semblable à l'analyse de toute autre langue morphologiquement pauvre. Les résultats vont différer en raison des ressources disponibles et de la propagation des erreurs venant des niveaux d'analyse précédents (segmentation et étiquetage morpho-syntaxique). Les algorithmes utilisés sont le plus souvent indépendants de la langue et procèdent par apprentissage automatique à partir d'un corpus arboré manuellement donné en entraînement. Des systèmes effectuant des analyses syntaxiques en constituants et en dépendances sont aujourd'hui disponibles. Le corpus arboré le plus utilisé dans les publications internationales est le Chinese Treebank (Xue et al. 2005).

D'autres corpus arborés sont disponibles, mais moins utilisés dans les publications internationales. Le Chinese Treebank est souvent préféré par souci de comparabilité des résultats, sa diffusion via le LDC (Linguistic Data Consortium basé aux États-Unis) plus aisée que les corpus chinois et taïwanais est sans doute un autre facteur important de sa popularité.

De récents travaux Zhao (2009), Zhang et al. (2013, 2014) visent à effectuer une analyse syntaxique directement à partir des caractères, sans séparer les tâches de segmentation, d'étiquetage et d'analyse syntaxique. Ces approches permettent d'éviter la propagation de certaines erreurs de segmentations et améliorent ainsi les résultats finaux. Cependant ces tâches jointes ne résolvent pas la question du «mot» en chinois. Elles reformulent et déplacent le problème d'un découpage d'une séquence linéaire à un étiquetage des relations de dépendances entre deux caractères (qui sont qualifiées de *intra-* ou *inter-mot* dans Zhang et al. (2014). Si elle ne résout pas le problème linguistique sous-jacent, l'approche jointe permet tout de même une amélioration des systèmes d'analyse automatique et une reformulation intéressante de la question.

6. NIVEAUX SEMANTIQUES ET DISCURSIFS

6.1 Étiquetage en rôles sémantiques

[12] La catégorie syntaxique la plus polysémique.

Le premier niveau d'analyse sémantique (au-delà de la désambiguïsation lexicale présentée en section 3.4) consiste en l'identification et l'analyse des relations entre prédicats et arguments, une tâche appelée *étiquetage en rôles sémantiques (Semantic Role Labelling)*. Le corpus d'apprentissage le plus utilisé pour cette tâche est le PROPBANK (*Proposition Bank*) du mandarin (Xue 2008). Les systèmes d'étiquetage pour le mandarin sont globalement meilleurs que ceux appris pour l'anglais alors même que les jeux de données disponibles sont moins conséquents. Une syntaxe plus uniforme concernant en particulier l'ordre des adjoints rend l'appariement des rôles sémantiques et des constituants syntaxiques de surface plus aisé qu'en anglais par exemple. Cependant, cette analyse dans un scénario réaliste pour le mandarin est encore un défi étant donné qu'il requiert la segmentation, l'étiquetage morpho-syntaxique et une analyse syntaxique de surface comme point d'entrée. Il n'est donc pas interdit de supposer qu'une partie plus importante, en mandarin qu'en anglais ou français, des questions sémantiques soient résolues dans ces trois phases antérieures à l'étiquetage en rôles sémantiques.

6.2 Résolution anaphorique

En TAL, les tâches de résolution anaphorique et d'établissement de la coréférence sont distinguées. La première consiste essentiellement à retrouver dans le contexte discursif, l'expression linguistique à laquelle une expression anaphorique (pronom, défini ou autre) est associée. La deuxième vise à établir les chaînes de coréférence d'un texte (voire en dehors du texte) et ce quelle que soit la nature des expressions introductrices (pronoms, définis, mais aussi noms propres). Ces tâches sont cruciales pour la plupart des applications demandant une extraction d'informations de haut niveau. En mandarin, le travail s'est focalisé sur un problème particulier et relativement nouveau dans ce contexte: la résolution des pronoms vides (*zero-anaphora resolution*). En effet, le mandarin autorise, en particulier, l'omission du sujet quand celui est donné discursivement. D'autres langues, comme l'espagnol, permettent également cela, mais dans ce cas, les marques morphologiques sur le verbe permettent au moins d'identifier partiellement le pronom omis. En mandarin, ce n'est pas le cas et la résolution de ces pronoms vides, très fréquents, est un réel défi pour le TAL.

Yeh et al. (2007) ont proposé un modèle symbolique utilisant les prédictions de théories du discours comme *la théorie du centrage* en réalisant une analyse syntaxique peu profonde afin d'obtenir les informations syntaxiques nécessaires à l'application de cette théorie. Ce sont cependant les méthodes d'apprentissage automatique qui dominent la scène pour cette tâche aussi. Deux familles de méthodes ont émergé : celles utilisant simplement des listes de traits (provenant de la phrase en cours de traitement, de la phrase précédente ou du contexte plus général) et celles exploitant la structure locale Kong & Zhou (2010) et qui extraient systématiquement un peu de structure des exemples vus afin de construire un modèle.

Très récemment une approche non-supervisée Chen & Ng (2014) est parvenue à rivaliser avec l'état de l'art en méthode supervisée. Cette méthode utilise un résolveur probabiliste pour apprendre un modèle sur la base des exemples où le pronom est présent puis applique le modèle appris dans les cas où il est absent.

6.3 Analyse du discours

L'analyse automatique de la structure discursive est un défi quelle que soit la langue considérée. Cette tâche n'est ni plus ni moins qu'une tentative de modéliser la cohérence des textes. Le travail systématique d'analyse manuelle de ces structures a démarré pour le mandarin (Zhou & Xue 2012) en appliquant les principes du Penn Discourse Treebank (Prasad et al. 2008). La plupart des projets et des approches se sont focalisés sur l'utilisation d'indices explicites comme les marqueurs de discours signalant des relations spécifiques (temporelles, causales, argumentatives...). Les travaux sur le mandarin débutent sur ces questions et présentent deux particularités: (i) une ambiguïté encore plus sévère que pour l'anglais ou le français de ces marqueurs pris individuellement, (ii) leur fonctionnement très régulier en paire qui limite un peu cette ambiguïté. Le travail de Huang et al. (2014) vise la désambiguïsation des fonctions discursives de marqueurs comme 只不過/*zhǐ búguò*, notamment en étudiant statistiquement l'appariement de paires de ces marqueurs comme 雖然 … 但是 *suīrán … dànshì*. Les résultats sont très variables d'un marqueur à l'autre, mais le système propose généralement une fonction discursive préférée pour chaque marqueur de surface.

7. RESSOURCES LINGUISTIQUES

7.1 Lexiques

Le lexique du CKIP (Chinese Knowldege Information Processing) [13] développé à l'Academia Sinica a été le premier lexique électronique incluant des informations syntaxiques. Il a été utilisé intensivement pour le traitement automatique des langues à Taïwan en particulier. Le GKBCC (Grammatical Knowledge Base of Contemporary Chinese) de Yu et al. (1998) inclut la prononciation, la catégorie syntaxique, le sens et des informations sur l'usage grammatical des mots et idiomes du mandarin.

Dans le paysage des lexiques, Wordnet (créé initialement pour l'anglais seulement) (Fellbaum 1998) a une place particulière. Il s'agit en réalité d'un thésaurus électronique compilé initialement pour les besoins d'expériences psycholinguistiques. Néanmoins son utilité dans un grand nombre d'applications a conduit à atteindre une grande qualité et une grande couverture pour cette

[13] http://www.aclclp.org.tw/use_ckip.php

ressource. Wordnet est organisé sémantiquement autour d'ensembles de synonymes (*synsets*) et de relations entre *synsets* ou entre mots. C'est par conséquent une réelle opérationalisation d'un lexique relationnel où le sens de chaque mot est en large partie déterminé par les relations qu'il entretient avec les autres mots. Il existe maintenant plusieurs *wordnets* du mandarin comme par exemple le CHINESE WORDNET (Huang et al. 2010) développé à Taïwan, qui est librement consultable.[14] Une visualisation originale du graphe construit à partir des liens de synonymie du CWN est disponible à l'adresse http://naviprox.net/tmuse/ (Chudy et al. 2013). Enfin, il faut noter l'initiative de liage des différents wordnets disponibles (impliquant en particulier le mandarin) (Bond & Foster 2013 ; Wang & Bond 2013).

L'association de ces lexiques avec des informations de nature plus encyclopédique est cruciale pour un grand nombre d'applications TAL (Prévot et al. 2010). La méthode dominante a longtemps été d'associer un thésaurus électronique du type de ceux présentés ci-dessus à des ontologies (des spécifications formelles de connaissances encyclopédiques ou de sens commun). Les travaux sur le mandarin sont pionniers dans ce domaine avec notamment Sinica-BOW (Huang et al. 2004) qui associe le WordNet anglophone à des lexies du mandarin (via des liens de traduction) puis à SUMO, une ontologie générique (Niles & Pease 2003). La version du WordNet Mandarin utilisé dans cette ressource inclut près de 100 000 ensembles de synonymes.

7.2 Corpus annotés

Cela a été entrevu dans les sections précédentes, les corpus, en particulier ceux qui sont annotés, sont devenus le nerf de la guerre en traitement automatique des langues. En ce qui concerne le mandarin, de grands corpus annotés morpho-syntaxiquement ont été produits. Chaque corpus suit les standards en cours dans la zone dans lequel il est publié. Par exemple, le «Corpus Equilibré de mandarin Moderne de l'Academia Sinica» (Chen et al. 1996) inclut plus de 10 millions de mots issus de sources et de genre variés. Le corpus Gigaword (Graff et Chen 2005 ; Ma et Huang 2006) développé par le LDC (Linguistic Data Consortium) et annoté lui aussi à l'Academia Sinica inclut 1,4 milliard de caractères provenant de sources de presses de Taïwan, de Chine continentale et de Singapour. La segmentation et l'étiquetage ont été effectués automatiquement et vérifiés partiellement. Sa précision est au-delà de 95 %.

Le LIVAC (Linguistic Variations in Chinese Speech Communities, Hong Kong) (Tsou et al. 1997) est un corpus de suivi (*monitor corpus*) échantillonnant régulièrement un nombre important de sources de presse provenant des grandes villes de Chine continentale ainsi que de Hong Kong, Macau, Taïwan et Singapour. De 1995 à 2013, 500 millions de caractères ont été ainsi traités dans

[14] http://lope.linguistics.ntu.edu.tw/cwn/

ce cadre. Des travaux portant en particulier sur les noms propres en mandarin ont pu être développés sur cette base.

Le *Sinica Treebank* (Chen et al. 1999 ; Chen et al. 2003) inclut une analyse des structures syntaxiques en constituants et rôles sémantiques d'une portion du Sinica Corpus. Le schéma d'annotation suit une «*Information Case Grammar*». Cette double annotation peut permettre d'obtenir un corpus annoté en dépendances marquées par des rôles sémantiques, mais est en pratique peu utilisé en TAL du mandarin.

Le *Chinese Treebank* (Xue et al. 2005) est à ce jour le corpus le plus utilisé. Sa compilation a débuté en 1998 à l'université de Pennsylvanie. Il est aujourd'hui maintenu et enrichi à l'université de Brandeis et disponible en version 8.0 via le *Linguistic Data Consortium*[15]. Il compte maintenant 1,6 million de mots d'origines diverses, mais principalement des textes de presse (il inclut des articles de journaux, magasines, transcriptions d'émissions et blogs). Son annotation est largement inspirée du Penn Treebank pour l'anglais. Elle comporte segmentation, étiquetage en parties du discours et analyse syntaxique. L'annotation syntaxique a commencé en constituants puis des fonctions syntaxiques ont été ajoutées. Elles permettent une conversion en dépendance en utilisant des règles de recherche des têtes de syntagmes. Le résultat de l'application de ces règles n'a pas, à notre connaissance fait l'objet de correction manuelle ou d'évaluation comme cela a pu être le cas pour le passage en dépendance du French Treebank par (Candito et al. 2009). En outre, la comparaison de deux méthodes de conversion différentes a révélé un taux d'accord très faible (Xue 2007). On peut donc mettre en doute la qualité linguistique de ces annotations obtenues semi-automatiquement. Nous regrettons aussi le peu de documentation concernant les relations de dépendances utilisées, qui sont principalement calquées sur l'anglais. À l'inverse, les deux corpus qui suivent ont été dès le départ annotés manuellement en dépendances. L'annotation du *Peking University Multi-view Chinese Treebank* (PMT) a même été pensée pour que la conversion automatique d'une structure de dépendance en une structure en constituants soit fiable.

Le *Chinese Dependency Treebank* (Liu et al. 2006)[16] développé au *Harbin Institute of Technology Research Center for Social Computing and Information Retrieval* est un corpus segmenté, étiqueté et analysé syntaxiquement en dépendances de près d'un million de mots.

Le *Peking University Multi-view Chinese Treebank* (PMT) est un nouveau corpus proposé par l'Institut de Linguistique Computationelle de l'Université de Pékin (Qiu et al. 2014). Ce corpus est disponible gratuitement et compte 14 463 phrases. Chaque phrase est segmentée en mots qui sont annotés en parties du

[15] Référence LDC2013T21.
[16] Référence LDC2012T05.

discours (en utilisant un jeu de 11 étiquettes qui forment une version simplifiée du jeu d'étiquettes du corpus de PKU) et en dépendance (en utilisant un jeu de 32 relations). Un programme de conversion pour obtenir une annotation en constituants est disponible[17].

8. OUTILS LIBRES ET APPLICATIONS

Étant donné la place grandissante du mandarin dans le TAL, un nombre important de logiciels d'analyse ont été proposés. Parmi ceux-ci certains sont librement téléchargeables et gratuitement utilisables. Ils peuvent ainsi être utilisés comme outils pour la recherche en linguistique chinoise ou autres disciplines nécessitant l'analyse de textes en mandarin. Nous proposons ici un panorama rapide des outils librement disponibles que chacun peut utiliser et adapter à ses besoins. Il est important de constater que les différentes applications liées au TAL peuvent donner lieu à des adaptations spécifiques concernant les niveaux de traitement introduits ci-dessus. C'est par exemple le cas de Zeng et al. (2014) qui adaptent leur segmentation dans la perspective de la traduction automatique. Leur méthode consiste à intégrer dans leur modèle statistique de segmentation des informations sur l'alignement des séquences de caractère avec la langue cible. La prise en compte des spécificités des textes est elle aussi un domaine actif avec par exemple de travail de Li & Xue (2014) qui inclut des informations provenant du document pour une tâche de segmentation dans des textes de brevets.

8.1 Concordanciers en ligne

Certains corpus, annotés manuellement ou pré-analysés sont disponibles directement en ligne pour des recherches de type «concordance». Ils présentent l'avantage d'être simples à utiliser, car ils ne nécessitent aucune installation et utilisation de logiciel d'analyse, mais c'est au prix d'une certaine absence de souplesse. L'utilisateur ne peut pas changer la constitution du corpus et les statistiques qu'il peut obtenir sont celles définies *a priori* par les auteurs du site.

Interface vers le Sinica Corpus : http://app.sinica.edu.tw/kiwi/mkiwi/
Interface vers les corpus de l'université de Pékin : http://ccl.pku.edu.cn/corpus.asp
Concordances hébergées à l'université de Leeds : http://corpus.leeds.ac.uk/query-zh.html

8.2 Analyseurs morpho-syntaxiques

[17] On peut l'obtenir en contactant les auteurs via le site http://klcl.pku.edu.cn/ShowNews.aspx?id=137.

Dans cette section, nous signalons quelques projets Open Source qui permettent d'effectuer les différentes tâches décrites ci-dessus. Ils peuvent être utilisés comme boîtes à outils pour la linguistique de corpus ou servir de base à de futurs travaux en traitement automatique des langues. Ces logiciels sont librement et gratuitement téléchargeables et modifiables.

ZPar
ZPar est le système proposé par Zhang et al. (2014) et ses travaux précédents. Il permet d'effectuer segmentation, étiquetage en parties du discours et analyse syntaxique en constituants comme dépendances.[18]

Stanford NLP
Le groupe de traitement automatique des langues de Stanford propose un ensemble d'outils de TAL qui incluent segmentation, étiquetage en parties du discours, reconnaissance d'entités nommées, analyses syntaxiques et résolution de coréférences[19].

Berkeley
Le «Berkeley NLP Group» propose son analyseur syntaxique en constituants au libre téléchargement (licence GPLv2). Un modèle pré-entraîné sur le *Chinese Treebank* est disponible.

9. CONCLUSION

Ce panorama aura permis d'introduire la situation du traitement automatique du mandarin à la communauté des linguistes travaillant cette langue. Il est excitant de noter que cet état de l'art évolue très vite.

Il est en revanche presque surprenant de constater que globalement la définition des tâches et des techniques fondamentales du TAL n'a pas évolué radicalement ces dernières années. L'essentiel des travaux contemporains porte sur l'augmentation de la couverture des systèmes, en particulier pour traiter des nouvelles formes de langues très représentées sur internet et dans les moyens de communication modernes. Nous pouvons néanmoins citer quelques avancées qui, sans modifier les paradigmes principaux, constituent des évolutions notables dans le domaine. Par exemple, la sémantique prend un rôle plus important qu'auparavant, notamment via l'interface entre le TAL et l'ingénierie des connaissances (transformations des informations extraites en de réelles bases de données structurées par exemple). D'autre part, les langues dites « peu dotées » (en ressources et en outils automatiques) sont plus systématiquement abordées dans cette perspective de traitement automatique.

Finalement, nous espérons que cet état des lieux donnera envie aux linguistes et aux spécialistes du TAL d'initier plus de collaborations, car comme nous

[18] http://www.sutd.edu.sg/cmsresource/faculty/yuezhang/zpar.html
[19] http://nlp.stanford.edu/software/index.shtml

l'avons souligné à plusieurs reprises, les questions et les fondations linguistiques sont encore trop peu présentes en TAL du mandarin. Plus grave, l'historique de la discipline conduit à privilégier la solution de facilité consistant à adapter des outils et méthodes éprouvés pour l'anglais au cas du mandarin. Les problèmes spécifiques (comme la question de la relation caractère - mot) que nous avons présentés montrent que parfois un changement de paradigme informé par les connaissances linguistiques serait sans doute préférable. Ce n'est pas nécessairement le meilleur choix d'efficacité à court terme, mais aborder ces problèmes de manière spécifique permettrait des solutions bien mieux adaptées à cette langue. Il n'est alors pas interdit de penser que c'est également le meilleur choix pour améliorer les performances sur le long terme.

REMERCIEMENTS

Ce travail a pu être réalisé grâce aux mobilités des deux premiers auteurs effectuées dans le cadre des projets financés par l'union européenne Erasmus Mundus Action 2 «Multilingualism and Multiculturalism» 2009-5259-5 et 2010-5094-7. Il a également bénéficié d'une aide du gouvernement français, gérée par l'Agence Nationale de la Recherche au titre du projet Investissements d'Avenir A*MIDEX portant la référence n°ANR-11-IDEX-0001-02. Les auteurs tiennent à remercier un relecteur anonyme pour ses commentaires précis, qui ont grandement contribué à améliorer l'article.

REFERENCES

Allen J. D., 2007, The unicode standard version 7.0 – core specification (7e éd.), consulté sur http://www.unicode.org/versions/Unicode7.0.0/ch16.pdf.

Bird S., 2006, NLTK : the natural language toolkit, in *Proceedings of the* COLING/ACL *Interactive Presentation Sessions*, Sydney, Australia, p. 69-72.

Bishop T. & Cook R., 2003, A specification for CDL, Character Description Language, Technical report, La Jolla, Ca., Wenlin Institute.

Boitet C., 2008, Les architectures linguistiques et computationnelles en traduction automatique sont indépendantes, in *Actes de la Conférence sur le Traitement Automatique des Langues naturelles* (TALN), Avignon, France.

Bond F. & Foster R., 2013, Linking and extending an open multilingual wordnet, in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics,* Sofia, Bulgaria, p. 1352–1362.

Candito M., Crabbé B., Denis P. & Guérin F., 2009, Analyse syntaxique du français : des constituants aux dépendances, in *Actes de la 16e Conférence sur le Traitement Automatique des Langues naturelles* (TALN)*,* Senlis, France.

Chen Chen & Ng Vincent, 2014, Chinese Zero Pronoun Resolution : An Unsupervised Probabilistic Model Rivaling Supervised Resolvers, in

Proceedings of the 2014 Conference on Empirical Methods in Nnatural Language Processing, Doha, Qatar, p. 763–774.

Chen Feng-Yi, Tsai Pi-Feng, Chen Keh-Jiann & Huang Chu-Ren, 1999, Sinica treebank, in *Computational Linguistics and Chinese Language Processing*, volume 4(2), p. 87-104.

Chen Keh-Jiann, Huang Chu-Ren, Chang Li-Ping & Hsu Hui-Li, 1996, Sinica corpus : Design methodology for balanced corpora, in *Proceedings of the 11th Pacific Asia Conference on Language, Information and Computation*, Seoul, Korea, p. 167-176.

Chen Keh-Jiann, Luo Chi-Ching, Chang Ming-Chung, Chen Feng-Yi, Chen Chao-Jan., Huang Chu-Ren. & Gao, Zhao-Ming, 2003, Sinica Treebank, in A. Abeillé, *Treebanks*, Volume 20, Springer, Netherlands, p. 231-248.

Chudy Y., Desalle Y., Gaillard B., Gaume B., Magistry P. & Navarro E., 2013, Tmuse : Lexical network exploration, in *Proceedings of the 6th International Joint Conference on Natural Language Processing*, Nagoya, Japan, p. 41-44.

Dang Hoa Trang, Chia Ching-yi, Palmer M. & Chiou Fu-Dong, 2002, Simple features for Chinese Word Sense Disambiguation, in *Proceedings of the 19th International Conference on Computational Linguistics*, volume 1, Philadelphia, USA, p. 1– 7.

Fellbaum C., 1998, *WordNet*. An electronic lexical database, Cambridge, Ma, MIT Press.

Graff D. & Chen Ke, 2005, Chinese gigaword, LDC Catalog No.: LDC2003T09, ISBN 1, 58563–58230.

Habert B., 2004, Outiller la linguistique : de l'emprunt de techniques aux rencontres de savoirs, *Revue française de linguistique appliquée*, 9(1), p. 5-24.

Heiden S., Magué J.-P., Pincemin B., et al., 2010, TXM : Une plateforme logicielle opensource pour la textométrie-conception et développement, in *Proceedings of 10th international Conference on the Statistical Analysis of Textual Data* - JADT *2010,* Roma, Italy, volume 2, p. 1021–1032.

Huang Chu-Ren, Chang Ru-Yng & Lee Hsiang-pin, 2004, Sinica BOW (Bilingual Ontological Xordnet): Integration of bilingual Wordnet and Sumo, in *Proceedings of the International Conference on Language Resources and Evaluation* (LREC), Lisbon, p. 1553-1556.

Huang Hen-Hsen, Chang Tai-Wei, Chen Huan-Yuan & Chen Hsin-Hsi, 2014, Interpretation of Chinese Discourse Connectives for Explicit Discourse Relation Recognition, in *Proceedings of the 25th International Conference on Computational Linguistics : Technical Papers*, Dublin, p. 632–643.

Huang Chu-Ren, Hsieh Shu-Kai, Hong Jia-Fei, Chen Yun-Zhu, Su I-Li, Chen Yong-Xiang & Huang Sheng-Wei, 2010, Chinese WordNet : design, implementation, and application of an infrastructure for cross-lingual knowledge processing, in *Journal of Chinese Information Processing*, volume 24(2), p. 14–23.

Jurafsky D. & James H., 2000, *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech,* Upper Saddle River, USA, Prentice Hall.

Kilgarriff A. & Rosenzweig J., 2000, Framework and results for English SENSEVAL, in *Computers and the Humanities*, volume 34(1-2), p. 15–48.

Kong Fang & Zhou Guodong, 2010, A tree kernel-based unified framework for Chinese zero anaphora resolution, in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Honolulu, Hawaii, p. 882–891.

Li Si & Xue Nianwen, 2014, Effective Document-Level Features for Chinese Patent Word Segmentation, in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, Baltimore, p. 199–205

Liu Ting, Ma Jinshan & Li Sheng, 2006, Building a dependency treebank for improving Chinese parser, in *Journal of Chinese Language and Computing*, volume 16(4), p. 207–224.

Ma Wei-Yun & Huang Chu-Ren, 2006, Uniform and effective tagging of a heterogeneous gigaword corpus, in *Proceedings of the 5th International Conference on Language Resources and Evaluation,* Genoa, Italy, p. 24–28.

Magistry P., 2013, Unsupervised Word Segmentation and Wordhood Assessment : The case for Mandarin Chinese, Thèse de doctorat non publiée, Université Paris Diderot, France.

Manning C. D. & Schütze H., 1999, *Foundations of statistical natural language processing*, Cambrige, Ma., MIT Press.

Niles I. & Pease A., 2003, Mapping WordNet to the Sumo ontology, in *Proceedings of the IEEE International Knowledge Engineering Conference*, Las Vegas, USA, p. 23–26

Prasad R., Dinesh N., Lee A., Miltsakaki E., Robaldo L., Joshi A. K. & Webber B. L., 2008, The Penn Discourse Treebank 2.0, in *Proceedings of the 6th International Conference on Language Resources and Evaluation* (LREC), Marrakech, Morocco.

Prévot L., Huang Chu-Ren, Calzolari N., Gangemi A., Lenci A. & Oltramari A., 2010, Ontology and the lexicon : a multidisciplinary perspective, in *Ontology and the lexicon : A natural language processing perspective*, Cambridge, Cambridge University Press, p. 3–24.

Qiu Likun, Zhang Yue, Jin Peng & Wang Houfeng, 2014, Multi-view Chinese Treebanking, in *Proceedings of the 25th International Conference on Computational Linguistics : Technical Papers*, Dublin, Ireland, p. 257–268.

Raghunathan K., Lee H., Rangarajan S., Chambers N., Surdeanu M., Jurafsky D. & Manning C., 2010, A multi-pass sieve for coreference resolution, in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Cambridge, Ma., USA, p. 492–501.

Sagart L., 2006, L'emploi des phonétiques dans l'écriture chinoise, in Bottero F. & Djamouri R., *Ecriture chinoise. Données, usages et représentations*, Paris, EHESS, p. 35-53.

Sagot B. & Boullier P., 2008, Sxpipe 2 : architecture pour le traitement pré-syntaxique de corpus bruts, *Traitement Automatique des Langues*, volume 49(2), p. 155–188.

Tanguy L. & Hathout N., 2007, *Perl pour les linguistes : programmes en Perl pour exploiter les données langagières*, Paris, Hermès science publications/ Lavoisier.

Tellier I., 2009, Apprentissage automatique pour le TAL : Préface, *Traitement Automatique des Langues*, volume 50(3), p. 7–21.

Tsou, B., Lin, Hing-Long, Chan, Terence, Hu, Jerome, Chew, Ching-Hai & Tse, John K. P., 1997, A Synchronous Chinese Language Corpus from Different Speech Communities: Construction and Application, *International Journal of Computational Linguistics and Chinese Language Processing*, volume 2(1), p. 91-104.

Wang Shan & Bond F., 2013, Building the Chinese Open Wordnet (COW): Starting from core synsets, in *Proceedings of the 11th Workshop on Asian Language Resources*, a workshop at IJCNLP, Nagoya, Japan, p. 10–18.

Wong Kam-Fei, Li Wenjie, Xu Ruifeng & Zhang Zheng-sheng, 2009, Introduction to Chinese natural language processing, *Synthesis Lectures on Human Language Technologies*, 2(1), p 1–148.

Wu Andy, 2003, Chinese Word Segmentation in MSR-NLP, in *Proceedings of the Second Sighan Workshop on Chinese Language Processing*, Sapporo, Japan, p. 172–175.

Xue Nianwen, 2007, Tapping the implicit information for the PS to DS conversion of the Chinese Treebank, in *Treebanks and linguistic Theories*, Bergen, Norway, p. 189– 200.

Xue Nianwen, 2008, Labeling Chinese predicates with semantic roles, *Computational Linguistics*, volume 34 (2), p. 225–255.

Xue Nianwen et al., 2003, Chinese word segmentation as character tagging, *Computational Linguistics and Chinese Language Processing*, volume 8(1), p. 29–48.

Xue Nianwen, Xia Fei, Chiou Fu-Dong & Palmer M., 2005, The Penn Chinese TreeBank : Phrase structure annotation of a large corpus, *Natural Language Engineering*, volume 11(02), p. 207–238.

Yeh Ching-Long & Chen Yi-Chun, 2007, Zero Anaphora Resolution in Chinese with Shallow Parsing, in *Journal of Chinese Language and Computing*, volume 17 (1), p. 41–56.

Yu Liang-Chih, Lee Lung-Hao, Tseng Yuen-Hsien. & Chen Hsin-Hsi, 2014, Overview of SIGHAN 2014 Bake-off for Chinese Spelling Check, in *Proceedings of the Third Cips-Sighan Joint Conference on Chinese Language Processing*, Wuhan, China, p. 126–132.

Yu Shiwen, Zhu Xuefeng & Wang Hui, 1998, *The grammatical knowledge-base of contemporary Chinese—a complete specification*, Beijing, Tsinghua University Press.

Zeng Xiaodong, Chao Lidia S., Wong Derek F., Trancoso I. & Tian Liang, 2014, Toward Better Chinese Word Segmentation for SMT via Bilingual Constraints, in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, Baltimore, USA, p. 1360–1369.

Zhang Hua-Ping, Yu Hong-Kui, Xiong De-Yi & Liu Qun, 2003, HHMM-based Chinese lexical analyzer ICTCLAS, in *Proceedings of the Second Sighan workshop on Chinese Language Processing,* Sapporo, Japan, p. 184–187.

Zhang Meishan, Zhang Yue, Che Wanxiang & Liu Ting, 2013, Chinese Parsing Exploiting Characters, in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, volume 1, Sofia, Bulgaria, p. 125–134

Zhang Meishan, Zhang Yue, Che Wanxiang & Liu Ting, 2014, Character-Level Chinese Dependency Parsing, in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, Baltimore, USA, p. 1326–1336.

Zhao Hai, 2009, Character-Level Dependencies in Chinese : Usefulness and Learning, in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, Athens, Greece, p. 879–887.

Zhou Yuping & Xue Nianwen, 2012, PDTB-style Discourse Annotation of Chinese Text, in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju Island, Korea, p. 69–77.

RESUME : Le mandarin, ou chinois standard est devenu le centre de toutes les attentions en traitement automatique des langues. Du point de vue scientifique, la quantité de données produites dans cette langue, notamment via les nouveaux moyens de communication, en fait un terrain de jeu privilégié pour le traitement automatique des langues (TAL) et en particulier l'apprentissage automatique. Dans cette article, nous passons en revue les principales tâches du TAL et signalons les difficultés spécifiques au traitement automatique du mandarin. Ces dernières sont intimement liées aux particularités linguistiques de cette langue. Nous espérons ainsi favoriser le dialogue entre les communautés des linguistes et des spécialistes du TAL du mandarin.

ABSTRACT : In recent years, Mandarin (or Modern Standard Chinese) has become the focus of a large body of works in Natural Language Processing (NLP). The large amount of data produced and readily available in this language through new communication channels create a fertile ground for NLP research, especially when relying on machine learning. In this article, we first introduce the NLP main tasks and methods and then give the state of the art in each task for Mandarin language. We underline the challenges that are specific to the processing of this language which are related to the specificity of Mandarin and of the Chinese script. We hope that this piece of research can and will enhance both dialogue and cooperation between the two communities of linguists and NLP practitioners working on Mandarin.