N°1
2025

Les Cahiers de l'AFPC

Association française des professeurs de chinois

La compréhension de l'écrit en classe de Chinois Langue Etrangère : Les apprenants de niveau élémentaire face à la segmentation des phrases en mots

Rachel Daveluy - Université de Montpellier Paul-Valéry – Laboratoire ReSO

Résumé

Contrairement à la majorité des langues écrites qui utilisent un espace blanc entre les mots pour les délimiter, la langue écrite chinoise, basée sur les caractères, ne marque pas de frontières claires entre les mots obligeant ainsi le lecteur à regrouper lui-même les caractères pour former des unités lexicales. Les études en psycholinguistique et en traitement automatique des langues ont montré que la segmentation des phrases en mots était un processus complexe aussi bien pour les lecteurs natifs que pour les apprenants de chinois langue étrangère. L'objectif de notre travail est de faire le lien entre non-visualisation du mot dans la phrase chinoise et compétence en compréhension de l'écrit (CE) chez des apprenants débutants de chinois. Nous avons proposé des tests de compréhension de l'écrit, des tests de segmentation et des questionnaires à des apprenants de chinois de niveau élémentaire pour connaitre leur point de vue sur l'absence de visualisation des mots, leur perception de leurs difficultés en CE et leur capacité à segmenter correctement des phrases chinoises en mots. Les résultats ont montré que les apprenants ont plus de facilité à identifier les mots rencontrés au tout début de leur apprentissage et qu'ils s'appuient sur les mots fonctionnels pour découper la chaine écrite. Les résultats ont également montré que la majorité des apprenants interrogés considère les caractères plutôt que les mots et qu'ils n'ont pas réellement conscience de la non-visualisation des mots ou des difficultés qu'elle génère. Pour autant, ils n'ont pas été capables de segmenter les phrases correctement. Les résultats de cette étude nous amènent à réfléchir à des pistes pour l'enseignement du chinois langue étrangère et notamment à l'introduction d'une démarcation entre les mots pour les apprenants débutants qui pourrait permettre une meilleure compréhension du message écrit chez ces derniers dans les premières phases d'apprentissage.

Mots-clefs : *Compréhension écrite du Chinois Langue Etrangère, segmentation des phrases en mots, apprenants débutants*

Introduction

De nos jours, la majorité des langues écrites ont recours à l'utilisation d'un espace blanc entre les mots, permettant au lecteur d'identifier rapidement les différentes unités lexicales. Néanmoins, la langue écrite chinoise ne marque pas de délimitation entre les mots. C'est également le cas d'autres langues écrites telles que le japonais, le thaï ou le tibétain. Un texte en chinois se présente sous la forme d'une chaîne de caractères de même taille et de

complexité variable, alignés les uns à la suite des autres et séparés chacun par un espace très mince (cette chaîne de caractères étant toutefois interrompue par des signes de ponctuation). Un caractère chinois correspond à une syllabe mais un mot équivaut à un ou plusieurs caractères. Il existe ainsi des mots monosyllabiques (d'un seul caractère), des mots dissyllabiques (de deux caractères) et des mots plurisyllabiques (de trois caractères ou plus). Puisque la longueur d'un mot chinois est variable et qu'il n'y a aucune démarcation entre les mots, le lecteur devra regrouper lui-même les caractères pour former des unités lexicales. Dès lors, « la lecture (qui est affaire de compréhension), comporte en chinois deux étapes : l'identification des caractères et la reconnaissance des mots » (Allanic, 2017, p.16).

Dans le domaine du traitement automatique du langage humain, la segmentation en mots des textes chinois est un défi majeur puisque les ordinateurs doivent regrouper les caractères en mots et que le repérage des unités lexicales est la toute première étape essentielle à tout autre processus de traitement automatique du mandarin (Prevot *et al.*, 2015). Les techniques de segmentation en linguistique computationnelle (fondées sur des dictionnaires et des données statistiques) se sont progressivement améliorées depuis le début des années 1980, la justesse de segmentation avoisinant aujourd'hui les 99% et la vitesse de segmentation les 100 000 mots par seconde (Mao, 2019). Dans le domaine de la psycholinguistique, l'étude de la vitesse de lecture et des mouvements des yeux (fixation et saccades) chez les Chinois natifs a permis de mieux comprendre le processus de segmentation en mots chez ces derniers et de s'apercevoir que, contrairement aux idées reçues sur cette langue non-alphabétique, le mot était saillant et avait bien une réalité psychologique pour les lecteurs chinois (Bai *et al.*, 2008).

Pour autant, l'absence de visualisation du mot dans la phrase chinoise est un sujet encore peu étudié dans le domaine de la didactique alors même que des pratiques de classes liées à la segmentation en mots pourraient très certainement améliorer la compréhension et la vitesse de lecture des apprenants de chinois de niveau élémentaire (Chen *et al.,* 2021 ; Shen *et al*, 2012). Ces chercheurs pensent qu'un marqueur entre les mots pourrait être utile aux apprenants de chinois, ce pourquoi, nous nous sommes interrogés sur le ressenti des apprenants. Nous avons mené une enquête auprès d'apprenants débutants pour connaître leur point de vue et aussi mieux cerner les difficultés qu'ils estiment rencontrer en compréhension de l'écrit (CE). Nous avons également voulu évaluer leur capacité à segmenter correctement des phrases en mots. Nous cherchons de cette façon à examiner la corrélation entre non-visualisation des mots chinois et compétence en CE du chinois langue étrangère.

1. Les spécificités de la langue écrite chinoise pouvant entraver la segmentation correcte des phrases en mots

1.1 La longueur des mots chinois n'est pas fixe

En chinois classique la majorité des mots sont monosyllabiques. Dans les *Entretiens de Confucius*[1], 91% des mots sont monosyllabiques, les autres mots étant pour la plupart des noms propres (Wang, 1978, cité par Ma, 2022). En revanche, à notre époque, la longueur d'un mot varie de 1 à 15 caractères (Liu *et al*., 2013) et la très grande majorité des mots sont

[1] Cf. Cheng A., *Entretiens de Confucius* (traduction intégrale du *Lunyu*, avec introduction, notes, cartes et chronologie), Paris, Éditions du Seuil, « Points-Sagesses », 1981, 180 p.

dissyllabiques. Les mots dissyllabiques représentent environ 70% de la totalité des mots, les mots monosyllabiques environ 15% à 20% et les mots plurisyllabiques environ 10% à 15% (Fan et Reilly, 2022 ; Liang *et al*., 2023 ; Shen *et al*., 2012). Pour autant, la fréquence d'utilisation des mots monosyllabiques est bien plus élevée que celle des autres mots, elle représente environ 60% de la fréquence totale d'utilisation des mots (Su, 2014, cité par Zhou et Zheng, 2020). Cette haute fréquence d'utilisation s'explique notamment par le fait que les mots monosyllabiques font majoritairement partie du lexique de base.

De plus, un même caractère peut soit constituer un mot à lui tout seul soit aussi faire partie d'un autre mot (combiné à un ou plusieurs autres caractères).[2] Huang et Xue (2012) donnent l'exemple des deux chaines de caractères [可 + 以] et [个 +人]. Après analyse d'un corpus de textes, ils ont découvert que la chaîne de caractère [可+以] apparaissait 879 fois en tant que mot dissyllabique *kěyǐ* (可以) « verbe pouvoir » et seulement 4 fois en tant que deux mots monosyllabiques 可 *kě* (可) « possibilité » et *yǐ* (以) dans le corpus étudié, ce mot avait le sens de « avec ». Cependant, dans le même corpus de textes, parmi les occurrences de la chaîne de caractères [个 +人], le mot dissyllabique *gèrén* (个人) « individu » apparaissait 143 fois et les deux mots monosyllabiques *gè* (个) « classificateur » et rén (人) « personne » apparaissaient 190 fois. De ce fait, le découpage correct d'une même chaîne de caractères est une tâche loin d'être aisée.

1.2 La capacité combinatoire des caractères et leur position dans les mots

Comme l'explique Allanic (2015), la capacité combinatoire des caractères est très élevée. Un même caractère peut être utilisé pour former un grand nombre de mots différents, notamment les caractères les plus usuels. A cela s'ajoute que la majorité des caractères peuvent apparaître à n'importe quelle position dans le mot et tout particulièrement les caractères les plus fréquents. Un même caractère peut de ce fait apparaître parfois en début, milieu ou fin de mot. Liang *et al.* (2017) estiment que seulement 20% des caractères communiquent des informations assez claires pour permettre une segmentation non-ambiguë de la phrase chinoise en mots.

Cependant, certains caractères apparaissent plus fréquemment à la même position dans les mots. Liang *et al* (2023) citent par exemple le caractère *hèn* (恨) qui apparaît une seule fois en début de mot parmi une liste de 17 mots. Le lecteur peut également identifier rapidement la frontière entre les mots en s'appuyant sur certaines catégories de mots comme les suffixes par exemple. Néanmoins, pour les caractères très usuels, la position du caractère dans le mot est loin d'être constante. Parmi 549 mots de quatre caractères pouvant être formés avec le caractère *rén* (人), ce dernier est placé 132 fois en première position, 161 fois en seconde, 113 fois en troisième et 143 fois en quatrième position (Li et Pollatsek, 2020).

[2] Un caractère *A* pourrait être le mot *A* mais aussi se retrouver dans les mots *AB*, *BAC*, *CAB, BCA, BCDA etc*.

Enfin, Drocourt (2022) souligne qu'un caractère représente une unité minimale de sens [3] et qu'un même caractère peut être utilisé pour noter plusieurs unités de sens très différentes. Par exemple, dans le Dictionnaire Ricci Chinois-Français (2014), le caractère *huā* (花) peut, entre autres, revêtir les significations suivantes : « fleur », « superficiel », « frivole », « courtisane », « flou », « dépenser de l'argent » ou « coton » et fait partie aussi de la liste des noms de famille.

1.3 L'identification ardue des noms propres, abréviations, néologismes et chengyu

La reconnaissance des noms propres est un problème majeur dans le traitement automatique du langage humain (Gu *et al.*, 2015 ; Liu *et al.*, 2022). Il n'y a pas de marqueur permettant d'identifier facilement un nom propre comme le ferait une lettre majuscule en français. Les noms chinois sont composés d'un nom de famille suivi d'un prénom, chacun pouvant être monosyllabique ou dissyllabique (les noms de famille dissyllabiques étant rares). Mais les transcriptions des noms des minorités ethniques ou des personnes étrangères n'ont pas de longueur fixe (Liu *et al.*, 2022). L'écrivain français Emile Zola sera transcrit en 埃米尔·左拉 et le compositeur allemand Jean-Sébastien Bach en 约翰·塞巴斯蒂安·巴赫, par exemple. Les cent noms de famille chinois les plus fréquemment utilisés représentent environ 87% de la population chinoise (Gu *et al.*, 2015) mais les prénoms peuvent être formés à partir de n'importe quel caractère. En outre, les noms de lieux prêtent souvent à confusion car ils peuvent déjà comprendre un autre nom propre tel qu'un nom de personne. Les noms d'organismes, quant à eux, sont souvent abrégés (Liu *et al.*, 2022). C'est le cas de l'Université de Pékin *Běijīng Dàxué* (北京大学) le plus souvent abrégée en *Běidà* (北大). Il est à noter que le lecteur pourra toutefois s'appuyer sur certains noms communs souvent utilisés avec les noms propres tels que « madame », « ville », ou « société » et ainsi repérer plus facilement ces noms propres. De plus, certains ouvrages soulignent les noms de personnes même si cette pratique est peu répandue si l'on prend en compte l'ensemble des corpus de textes existants.

La langue chinoise utilise beaucoup d'abréviations et la façon dont elles sont employées n'est pas toujours clairement définie. Ainsi le lecteur devra fortement s'appuyer sur ses connaissances préalables et le contexte pour découper correctement. Fu (2018) donne l'exemple de la chaine de caractères [中+非] qui selon les cas peut signifier le nom propre *Zhōngfēi* (中非) « La République centrafricaine », ou être une abréviation pour *Zhōngguó-Fēizhōu* (中国-非洲) « Chine-Afrique » et faire référence aux relations sino-africaines. A cela s'ajoute l'augmentation de l'emploi des abréviations dans les messageries instantanées et le langage Internet.

L'identification des néologismes est tout autant difficile puisqu'ils sont formés à partir de caractères préexistants et que toute chaîne de caractères pourrait potentiellement être un mot

[3] « En chinois moderne, [...] l'écrasante majorité des morphèmes chinois, soit plus de 93% d'entre eux, correspondent à une syllabe à l'oral et à un caractère à l'écrit. Ce qui revient à dire que, le plus souvent, le caractère représente une unité minimale de sens. » (Drocourt, 2022, p. 177-178)

nouveau. Les néologismes sont d'ailleurs, tout comme les abréviations, de plus en plus utilisés avec la propagation des outils numériques.

Les expressions idiomatiques telles que les *chéngyǔ* (成语) représentent aussi une réelle difficulté pour la segmentation des phrases en mots puisqu'elles sont le plus souvent composées de quatre caractères que seul un lecteur expérimenté sera capable de reconnaître en tant que telles.

En somme, la langue écrite chinoise ne marque pas clairement de frontières entre les mots et la longueur de ces derniers n'est pas fixe. La capacité combinatoire des caractères est très élevée et la position des caractères au sein des mots n'est pas constante. Les noms propres, abréviations, néologismes et *chéngyǔ* (成语) sont difficiles à identifier. C'est pour toutes ces raisons que toute chaîne de caractères chinois pourrait en définitive correspondre à un mot nouveau ou à un nom propre. De plus, de nombreux caractères peuvent avoir diverses significations et être associés de manières différentes aux caractères adjacents, ce qui fait que l'ambiguïté du message écrit reste une constante forte. « Lire et comprendre un message écrit consiste avant tout à 'découper' le continuum de la chaîne écrite en mots, compétence fondée non seulement sur la connaissance des caractères, mais aussi, sur celle du vocabulaire et de la syntaxe » (Drocourt, 2007, p.357). Ainsi, la segmentation appropriée de la phrase chinoise en mots repose fortement sur les diverses compétences du lecteur et le lecteur non aguerri pourra commettre des séries d'erreurs entravant grandement sa compréhension.

2. Le processus de segmentation des phrases en mots chez les lecteurs chinois natifs

Comme le soulignent Liu *et al.* (2013), les lecteurs chinois ne sont parfois pas d'accord sur la façon de découper les phrases en mots et ne placent pas toujours les frontières aux mêmes endroits dans la chaîne écrite. Ces chercheurs nous expliquent que la notion de mot est assez vague chez les lecteurs natifs qui ont tendance à combiner les mots monosyllabiques avec les mots dissyllabiques adjacents pour former une seule et même unité lexicale. Ils vont par exemple regrouper systématiquement le classificateur avec l'adjectif numéral qui le précède ou l'adjectif avec le nom qui le suit. En revanche, toujours selon Liu *et al.* (2013), dans certains cas les lecteurs natifs ont du mal à se mettre d'accord sur la façon de segmenter une phrase, par exemple, lorsque les adverbes sont suivis d'un adjectif, d'une préposition, d'un verbe ou d'un autre adverbe ou encore lorsque la phrase comporte une chaine consécutive de noms. Liu *et al.* (2013) pensent d'ailleurs que la tendance des lecteurs natifs à regrouper les caractères en morceaux de langue plutôt qu'en mots pourrait faciliter et améliorer leur compréhension en diminuant le nombre d'informations stockées en mémoire de travail.

Pour autant, le mot a bien une réalité psychologique pour les lecteurs chinois et les études ont montré que le mot est plus saillant pour eux que le caractère (Bai *et al.*, 2008). D'après Li *et al.* (2009) et Li et Pollatsek (2020), la segmentation et la reconnaissance des mots se fait en même temps chez les lecteurs chinois qui coordonnent les processus de haut niveau (connaissances antérieures, indices textuels) et de bas niveau (décodage linguistique). Et ces chercheurs estiment que c'est le processus de segmentation en mots qui influence en partie la reconnaissance des caractères. En outre, face à un mot inconnu les lecteurs natifs liront

les caractères un par un ce qui n'est pas le cas quand ils sont confrontés uniquement à des mots connus puisque qu'ils les identifient comme unités de lecture.

En plus de se référer au contexte et à leurs connaissances lexicales et syntaxiques pour réussir à segmenter correctement, les lecteurs chinois vont également s'appuyer sur la fréquence de position d'un caractère au sein des mots et plus particulièrement sur la probabilité d'être placé en fin de mot (Liang *et al.*, 2023).

3. L'impact de l'ajout d'une délimitation entre les mots pour la lecture du chinois

3.1 Chez les locuteurs natifs

De nombreux travaux ont démontré que l'ajout d'une démarcation entre les mots (espace blanc ou altération de couleurs) n'avait pas d'impact significatif sur les compétences en lecture des Chinois natifs (Bai *et al.*, 2008 ; Bassetti, 2009 ; Oralova et Kuperman, 2021) mais que l'ajout d'une délimitation entre chaque caractère diminuait leur vitesse de lecture et perturbait leur compréhension (Bai *et al.*, 2008).

Pour autant, l'ajout d'une frontière visible entre les mots permet aux enfants chinois d'acquérir plus facilement du vocabulaire nouveau (Blythe *et al.*, 2012) et aux adultes chinois d'améliorer la lecture de mots inconnus ou d'un lexique technique (Perea et Wang, 2017). Blythe *et al.* (2012) soulignent que lorsque l'on présente des mots nouveaux à des enfants chinois dans un texte espacé (espaces blancs entre chaque mot), ils arrivent beaucoup mieux à les identifier quand ils les revoient ensuite dans un texte non-espacé, en comparaison des enfants à qui on avait présenté les mots nouveaux dans un texte au format habituel non-espacé. Les résultats de l'étude de Blythe *et al.* (2012) nous forcent ainsi à nous interroger sur les bénéfices que pourrait apporter l'ajout d'une démarcation entre les mots pour la compréhension écrite du chinois chez les apprenants de niveau élémentaire.

3.2 Chez les apprenants de Chinois Langue Etrangère

Il n'y a pas à l'heure actuelle de réel consensus sur les bienfaits de l'ajout d'une délimitation entre les mots pendant la lecture chez les apprenants de chinois de niveau intermédiaire ou avancé (Chen, 2021). Cependant, l'ajout d'un espace blanc entre les mots facilite la lecture des apprenants de chinois de niveau élémentaire quelle que soit la langue écrite maternelle des apprenants. C'est à dire que l'effet est le même sur des apprenants de chinois japonais dont la langue maternelle ne fait pas apparaître d'espace entre les mots et les apprenants de chinois américains dont la langue maternelle fait apparaître des espaces (Bai *et al.*, ,2008 ; Shen *et al.*, 2012). En outre, l'altération de couleurs facilite la lecture du chinois chez les apprenants et cela tout en conservant le format habituel non-espacé d'un texte chinois (Zhou *et al.*, 2020)

Dans tous les cas, lorsque l'on compare la vitesse de lecture des apprenants de chinois dans un texte segmenté [aux mots] avec celle dans un texte segmenté [aux caractères], on s'aperçoit que la vitesse de lecture est beaucoup plus lente dans le deuxième cas, et cela aussi bien sur les apprenants de niveau avancé, qu'intermédiaire ou élémentaire (Shen *et al.*, 2012).

L'ajout d'espaces blancs ou de couleurs pour délimiter les mots dans les textes à destination des apprenants débutants pourrait leur être bénéfique en termes d'apprentissage de mots

nouveaux, leur permettre de lire plus vite et de lever l'ambiguïté du message écrit. La première étape essentielle pour comprendre est de repérer où sont les mots. Ainsi, une approche basée sur un format écrit dans lequel les mots seraient espacés pourrait être utilisée par les enseignants en même temps que les approches basées sur la maîtrise des connaissances syntaxiques, sémantiques et contextuelles (Chen *et al.*, 2021)

4. Etude du ressenti et des compétences de segmentation des apprenants de niveau élémentaire

4.1 Questions de recherche

Les études en psycholinguistique et en linguistique informatique ayant démontré que la segmentation des mots en chinois était un processus complexe aussi bien pour les lecteurs chinois natifs que pour les ordinateurs, nous avons voulu connaître le point de vue des apprenants débutants de chinois sur la question. Quelles sont, selon leur perception, leurs difficultés en compréhension de l'écrit ? Ces derniers considèrent-ils la non-visualisation des mots dans la phrase chinoise comme une difficulté en compréhension de l'écrit ? Comment ces apprenants procèdent-ils au découpage des mots dans la phrase chinoise ? Emploient-ils les mêmes stratégies que les lecteurs chinois natifs ? Leur façon de segmenter les phrases en mots pourrait-elle influencer leur capacité de compréhension ?

4.2 Les participants

Une enquête a été menée en mai 2024 auprès de 55 étudiants de chinois scolarisés en première année de licence de Langues Etrangères Appliquées (LEA). Les étudiants interrogés étaient tous francophones et avaient tous pour langue maternelle une langue écrite faisant apparaître un espace blanc entre les mots. Parmi les 55 étudiants ayant participé à cette étude, 50 étaient des grands débutants au mois de septembre 2023 et avaient donc suivi 192h de cours de langue chinoise sur 26 semaines (environ 7h par semaine) au moment de l'étude. Les 5 participants restants avaient déjà étudié le chinois au lycée ou en option à l'université. L'ensemble des participants possédaient un niveau de chinois élémentaire (niveau A1-vers A2 du CECRL) au moment de l'enquête. Il s'agit ici d'un échantillon tiré d'une enquête plus large et qui a été réalisée sur 169 apprenants de chinois en collège, lycée et université.

4.3 Déroulé de l'enquête

L'enquête s'est déroulée en classe, sous la surveillance d'un enseignant, dans le silence et sans aucun document autorisé. L'enquête a duré 45 minutes en moyenne.

L'enquête était divisée en quatre étapes distinctes. Chaque étape ayant été relevée par l'enseignant au fur et à mesure.

Dans une première étape, les apprenants ont réalisé une activité de compréhension de l'écrit (CE) à partir de documents authentiques (extraits de conversation de messagerie instantanée[4], extraits d'un article de journal, extraits d'une histoire pour enfants). Lors de cette première étape les apprenants ont dû traduire ou résumer en français le contenu des documents écrits.

[4] Extraits de conversations authentiques transmises par des sinophones.

Dans une deuxième étape, les participants ont répondu à deux questions ouvertes sur les difficultés qu'ils avaient pu rencontrer lors de l'activité de CE et ce sur quoi ils s'étaient appuyés pour comprendre les documents écrits.

Dans une troisième étape, les étudiants ont répondu à deux questions à choix multiples, une sur leurs difficultés en CE et une sur leurs stratégies.

Dans une dernière étape, les apprenants ont réalisé un exercice de segmentation : la consigne était de découper des phrases en mots à l'aide de barres obliques puis d'indiquer le nombre de mots pour chacune des phrases.

4.4 Choix des phrases à segmenter de l'étape 4 de l'enquête

Nous avons choisi de donner aux participants un total de huit phrases à segmenter. Parmi ces huit phrases, quatre étaient extraites des documents de CE de l'étape 1 et six étaient extraites d'articles de journaux en ligne. La longueur des phrases variait de 6 à 21 caractères et de 4 à 14 mots. Pour l'ensemble des phrases le nombre de caractères inconnus s'échelonnait de 0 à 5 et le nombre de mots inconnus de 0 à 3. Nous avons choisi ces phrases pour les raisons suivantes : les participants connaissaient la très grande majorité des caractères et des mots ; les phrases sélectionnées comportaient des mots inconnus construits à partir de caractères déjà connus tels que *fāngmiàn* (方面) « aspect » et *shūfǎ* (书法) « calligraphie » ; des noms de lieux tels que *Shēnzhèn* (深圳) la ville de Shenzhen et *Tiān'ānmén guǎngchǎng* (天安门广场) la place Tian'anmen ; un même mot *xuéxí* (学习) « étudier » placé différemment dans deux des phrases ; des prépositions, conjonctions, particules et suffixes pouvant servir de point d'appui pour la segmentation (tels que *zài* (在), *bǐ* (比), *hé* (和), *de* (得), *de* (的), *men* (们)) ; enchaînements de caractères pouvant induire des erreurs de découpage. [5]

5. Analyse synthétique des résultats d'un échantillon de l'enquête

5.1 Les apprenants débutants considèrent les caractères plutôt que les mots

Les apprenants de niveau élémentaire interrogés considèrent les caractères plutôt que les mots, ce qui, pour commencer, diffère de ce qui a été constaté chez les Chinois natifs dans les recherches citées plus haut. Si nous regardons les données collectées dans notre enquête, nous observons que les 55 participants interrogés ont plus souvent utilisé le terme « caractère » que le terme « mot ». Tout d'abord, dans les réponses données par les participants aux deux questions ouvertes de l'étape 2 (questions sur leurs difficultés en CE et les stratégies qu'ils utilisent), nous dénombrons 136 occurrences du terme « caractère », 36 occurrences du terme « mot », 10 occurrences du terme "vocabulaire » et 2 occurrences du terme « idéogramme ».

[5] Comme un enchaînement de caractères [ABCD] pouvant former différents mots [ABC + D] ou [AB+CD] à segmenter correctement selon le contexte et les informations syntaxiques. Dans la chaîne de caractères [大学生活] *dàxué shēnghuó* (la vie à l'université) le mot *dàxuéshēng* (大学生) « étudiant » existe mais c'est au mot *dàxué* (大学) « université » qu'il faut découper.

D'autre part, à la question à choix multiples de l'étape 3 « Quelles sont les difficultés que vous rencontrez quand vous devez comprendre des phrases ou des textes en chinois ? », les apprenants ont placé la réponse « les caractères inconnus » en première position des difficultés rencontrées. Ils devaient choisir jusqu'à cinq réponses en les classant par ordre d'importance en les numérotant de 1 à 5. Nous pouvons voir dans le tableau 1 que 33 participants sur 55 ont positionné la réponse « les caractères inconnus », en choix n°1 et qu'au total 53 participants (96,4%) ont choisi cette réponse comme une des cinq difficultés rencontrées. 41 participants (74.5%) ont choisi la réponse « mots inconnus » mais seulement 7 l'ont placée en premier choix des difficultés rencontrées.

Bai *et al.* (2008) ont démontré que l'ajout d'un espace blanc entre les caractères avait un impact négatif chez le lecteur en termes de vitesse et de compréhension. Ainsi, Il est fort probable que le fait que les apprenants débutants considèrent davantage les caractères que les mots puisse avoir une incidence sur leur rapidité de lecture et leurs capacités de compréhension.

	Choisi en n°1	Choisi en n°2	Choisi en n°3	Choisi en n°4	Choisi en n°5	non classé	Nombre total d'occurrences	total en %
Caractères inconnus	33	6	7	2	0	5	53	96.4%
Mots inconnus	7	23	6	0	2	3	41	74.5%
Différentes significations d'un mot	3	4	12	6	4	3	32	58.2%
Syntaxe/grammaire	2	11	9	5	2	1	30	54.5%
Regrouper les caractères en mots	0	5	4	6	3	1	19	34.5%
Abréviations	0	0	3	7	4	2	16	29.1%
Caractères qui se ressemblent	1	0	1	8	5	1	16	29.1%
Chengyu	3	1	2	3	2	2	13	23.6%
Noms propres	0	0	4	3	3	1	11	20%
Autre	1	0	0	0	1	0	2	3.6%

Tableau 1 : Les difficultés du point de vue des apprenants débutants

5.2 Les stratégies de segmentation des apprenants débutants

A l'étape 4 de l'enquête, les étudiants se sont livrés à un exercice de segmentation de phrases en mots. Après analyse des découpages proposés par les apprenants, nous remarquons tout d'abord qu'ils n'ont pas eu de difficultés à identifier les mots auxquels ils avaient été confrontés dès le début de leur apprentissage du chinois tel que le mot *Fǎguó* (法国) « France » qui a été correctement segmenté par 51 des participants (92,7 %) dans la phrase : « Le combien rentres-tu en France ? » (你几号回法国呀？).

De plus, ils ont été majoritairement capables de s'appuyer sur les mots fonctionnels pour découper correctement des chaînes de caractères comportant des mots ou caractères

inconnus. C'est le cas notamment du repérage correct de la conjonction de coordination *hé* (和) « et » qui a été identifiée par 45 participants (81,8 %) dans la chaîne de caractères [努力和认真] « travailler dur et être consciencieux » alors même qu'ils ne connaissaient ni les deux mots *nǔlì* (努力) « travailler dur » et *rènzhēn* (认真) « consciencieux », ni les caractères *nǔ* (努) et *rèn* (认). La majorité des participants (46 participants soit 83,6%) ont su également s'appuyer sur la préposition *zài* (在) dans la chaîne de caractères [在学习方面] « dans le domaine des études » pour délimiter correctement le mot *xuéxí* (学习) « étudier » alors que le mot *fāngmiàn* (方面) « aspect » leur est inconnu (même s'ils ont appris les deux caractères *fāng* (方) et *miàn* (面) dans d'autres mots). En contrepartie, ils n'ont majoritairement pas été capables de reconnaître le mot *xuéxí* (学习) « étudier » dans une autre des phrases proposées.

Dans le domaine du traitement automatique du langage humain, la distinction entre les unités syntaxiques et sémantiques est une étape essentielle au bon traitement des différentes unités linguistiques (Magistry, 2013). Ainsi, les apprenants semblent avoir ici procédé de la même façon en s'appuyant sur les mots fonctionnels.

5.3 Les erreurs de segmentation des apprenants débutants

Phrase 1 她说她那天晚饭吃得太多。 « Elle a dit qu'elle avait trop mangé l'autre soir au dîner. »	
Segmentation proposée	*Nombre d'occurrences*
晚饭 diner (segmentation correcte)	**34**
饭吃 (segmentation incorrecte)	11
晚饭吃 (segmentation incorrecte)	7
Phrase 2 大学生活是非常自由的, [...] « La vie à l'université, c'est la liberté, [...] »	
Segmentation proposée	*Nombre d'occurrences*
大学生 étudiant (segmentation incorrecte)	48
大学 / 生活 la vie à l'université (segmentation correcte)	**4**
大学 / 生 / 活 (segmentation incorrecte)	3
Phrase 3 在学习方面，北京大学得学生们非常努力和认真。 « Dans le domaine des études, les étudiants de l'université de Pékin sont très consciencieux et ils travaillent dur. »	
Segmentation proposée	*Nombre d'occurrences*
学习 (segmentation correcte)	46
Phrase 4 现在很多中国孩子从小学习书法。	

« De nos jours, beaucoup d'enfants chinois étudient la calligraphie depuis petits. »	
Segmentation proposée	*Nombre d'occurrences*
学习 (segmentation correcte)	**24**
小学 (segmentation incorrecte)	18
从小学 (segmentation incorrecte)	10

Tableau 2 : Exemples de segmentations proposées par les apprenants débutants

Il semblerait que les apprenants débutants aient tendance à découper la phrase au premier mot complet qu'ils aperçoivent lorsqu'ils sont confrontés à des mots inconnus sans forcément se préoccuper de la logique syntaxique ou sémantique de la phrase. En observant le tableau 2, nous nous rendons compte que moins de la moitié des participants (24 participants soit 43,6%) ont su repérer le mot *xuéxí* (学习) « étudier » dans la chaine de caractères [从小学习书法] « étudier la calligraphie depuis petit » alors qu'ils l'avaient bien repéré dans la chaine de caractères [在学习方面] « dans le domaine des études ». En effet, 18 participants ont associé les caractères *xiǎo* (小) et *xué* (学) pour former le mot *xiǎoxué* (小学) « école primaire » au lieu d'associer *xué* (学) à *xí* (习) pour faire le mot *xuéxí* (学习) « étudier ». Nous pensons donc qu'ils ne sont pas allés plus loin dans la phrase et ont découpé dès qu'ils ont vu ce mot. Ils ont commis le même type d'erreur avec la chaine de caractères [大学生活] « la vie à l'université » puisque 48 participants (87,3%) ont proposé ici le mot *dàxuéshēng* 大学生 « étudiant(e) » au lieu de proposer la segmentation [大学 + 生活]. Il faut néanmoins noter, dans ce cas, que les apprenants ne connaissaient pas le mot *shēnghuó* (生活) « vie ». Pour autant, les participants auraient pu proposer un découpage avec le mot *dàxué* (大学) « université » car ils connaissent ce mot et que le caractère *huó* (活) était suivi du verbe être *shì* (是) (cf. la phrase 3 dans le tableau 2). Cela rejoint les propos de Blythe *et al.* (2012) qui expliquent que la segmentation en mots est beaucoup plus difficile quand le lecteur est confronté à des mots nouveaux et qu'il ne sait pas comment regrouper les caractères. Au vu de ces résultats, il semblerait que l'absence de visualisation des mots ait été un frein à leur compréhension du message. En effet, même avec une recherche dans le dictionnaire ils n'auraient même pas pu retrouver les mots car ils n'ont pas regroupé les caractères correctement.

En outre, les apprenants débutants ont eu du mal à repérer le nom propre *Tiān'ānmén guǎngchǎng* (天安门广场) la place Tian'anmen. Seulement 23 participants (41,2%) ont su l'identifier et 17 participants (30,9%) ont regroupé t*iān* (天) avec *ān* (安) ce qui ne veut rien dire.

Quelques pistes en guise de conclusion

La langue écrite chinoise ne fait apparaitre aucune démarcation entre les mots ce qui ne cesse d'interroger les chercheurs en psycholinguistique sur le processus de lecture des lecteurs chinois natifs et qui oblige les chercheurs en traitement automatique des langues à faire appel à des algorithmes de plus en plus complexes. L'ajout d'une délimitation entre les mots (espace

blanc ou altération de couleurs) facilite la lecture de mots nouveaux chez les natifs et améliore la vitesse de lecture et les capacités de compréhension chez les apprenants de Chinois Langue Etrangère, notamment de niveau élémentaire. Nous avons donc interrogé des apprenants de chinois débutants pour connaitre leur point de vue sur l'absence de visualisation des mots, leur perception de leurs difficultés en compréhension de l'écrit et leur capacité à segmenter correctement les phrases chinoises en mots.

Quand nous comparons les données déclaratives des participants sur leur perception de leurs difficultés rencontrées en compréhension de l'écrit à leurs propositions de découpage lors de l'exercice de segmentation, il semblerait qu'une grande partie des apprenants débutants interrogés ne soit pas réellement consciente de la non-visualisation des mots ou du moins des difficultés de compréhension qu'elles génèrent. En effet, nous avons vu que les apprenants débutants considéraient davantage les caractères que les mots. De plus, nous pouvons voir dans le tableau 1 que 19 participants (34.54%) ont sélectionné la réponse « regrouper les caractères en mots » parmi la liste des difficultés proposées dans la question à choix multiple de l'étape 3 mais qu'aucun des apprenants ne l'a positionnée en premier choix. Pour autant nous avons vu, à travers les exemples donnés dans le tableau 2, que les apprenants n'avaient majoritairement pas segmenté correctement certaines phrases et que donc ils ne pourraient pas comprendre le message écrit.

Les résultats de cette étude nous amènent à réfléchir à des pistes pour l'enseignement-apprentissage du Chinois Langue Etrangère. Tout d'abord, la proposition faite par Shen *et al.* (2012) et Chen *et al.* (2021) de mettre des démarcations entre les mots dans les textes donnés aux apprenants de chinois parait tout à fait intéressante. Ce pourquoi, pour aller plus loin dans nos recherches, nous allons établir des tests de compréhension écrite et de lecture à haute voix de textes segmentés (faisant apparaître des espaces entre les mots) et de textes non segmentés à destination d'apprenants débutants de chinois en France. D'autre part, puisque les apprenants débutants ayant participé à l'enquête se sont appuyés sur les mots fonctionnels pour découper les phrases (comme le font les lecteurs chinois natifs et les algorithmes en traitement automatique du langage humain), il apparait assez évident qu'il ne faudrait pas trop négliger, en classe de chinois, l'enseignement-apprentissage de la grammaire. Enfin, « la maîtrise d'un nombre limité de caractères peut permettre la connaissance d'un nombre beaucoup plus grand de mots, surtout quand ce sont des caractères très usuels » (Allanic, 2015, p.4). De ce fait, le plus important ce n'est pas la taille du lexique enseigné aux apprenants de Chinois Langue Etrangère mais sa pertinence pour aider au mieux ces derniers à accéder plus efficacement et plus rapidement à la compréhension des documents écrits.

Références bibliographiques

ALLANIC Bernard, 2015, « Le débat sur la place attribuée aux caractères dans l'enseignement du chinois langue étrangère et l'émergence d'une école française de la disjonction oral/écrit » in Bouvier-Laffite Béatrice et Loiseau Yves (dir.), *Polyphonies franco-chinoise : mobilités, dynamiques identitaires et didactiques*, L'Harmattan, p. 143-158.

ALLANIC Bernard, 2017, *La voie des signes : l'apprentissage de la lecture en Chine*, Rennes, PUR.

BAI Xuejun et al., 2008, « Reading spaced and unspaced Chinese text: Evidence from eye movements » in *Journal of Experimental Psychology: Human Perception and Performance*, no 34, p. 1277-1287.
BASSETTI Bene, 2009, « Effects of adding interword spacing on Chinese reading: a comparison of Chinese native readers and English readers of Chinese as a second language » in *Applied Psycholinguistics, no 30* (4), p. 757-775.
BLYTHE Hazel I. et al., 2012, « Inserting spaces into Chinese text helps readers to learn new words: An eye movement study » *Journal of Memory and Language*, no 67(2), p. 241-254.
CHEN Ken et al., 2021, « Can Word-Word Space Facilitate L2 Chinese Reading: Evidence From the Two Empirical Studies by Advanced L2 Learners of Mandarin Chinese » in *Sage Open, no 11*(4), DOI:
https://doi.org/10.1177/21582440211059150 (consulté le 20/08/2024)
Dictionnaire Ricci Chinois-Français, 2014, Association Ricci,The Commercial Press.
FAN Xi, REILLY Ronan G., 2022, « Eye movement control in reading Chinese: A matter of strength of character? » in *Acta Psychol,* DOI: https://doi.org/10.1016/j.actpsy.2022.103711 (consulté le 17/09/2022)
FU Zhe, 2018, *Apprentissage non supervisé de la segmentation lexicale automatique du chinois basé sur les réseaux bayésiens avec application aux textes des médias sociaux*, Université du Québec à Montréal, Doctorat en informatique cognitive.
GU Chuan, TIAN Xi-ping, YU Jiang-de, 2015, « Automatic Recognition of Chinese Personal Name Using Conditional Random Fields and Knowledge Base » in *Hindawi Publishing Corporation Mathematical Problems in Engineering,* Volume 2015, DOI:
http://dx.doi.org/10.1155/2015/480879 (consulté le 05/09/2024)

HUANG Chu-Ren, XUE Nianwen 2012, « Words without Boundaries: Computational Approaches to Chinese Word Segmentation » in *Language and Linguistics Compass 6/8*, p. 494-505.
LI Xingshan, RAYNER Keith, CAVE Kyle R., 2009, « On the segmentation of Chinese words during reading » in *Cognitive Psychology*, no 58(4), p.5 25-552.
LI Xingshan, POLLATSEK Alexander, 2020, « An integrated model of word processing and eye-movement control during Chinese reading » in *Psychological Review,* Vol. 127, no 6,
p. 1139-1162
LIANG Feifei et al., 2017, « The role of character positional frequency on Chinese word learning during natural reading » in PLoS ONE 12(11): e0187656, DOI:
https://doi.org/10.1371/journal.pone.0187656 (consulté le 25/08/24)
LIANG Feifei et al., 2023, « The importance of the positional probability of word final (but not word initial) characters for word segmentation and identification in children and adults' natural Chinese reading » in *Journal of Experimental Psychology : Learning, Memory, and Cognition, no 49*(1), p. 98-115.
LIU Ping-Ping, 2013, « Do Chinese Readers Follow the National Standard Rules for Word Segmentation during Reading? » in *PLoS ONE 8(2)*:e55440. DOI:
0.1371/journal.pone.0055440 (consulté le 01/10/2022)
LIU Pan et al., 2022, « Chinese named entity recognition: The state of the art » in *Neurocomputing, no 473*, p. 37-53.
MA Chunyuan, 2022, *Ordre des mots et émergences du sens en français et en chinois*, Linguistique, Université Bourgogne Franche-Comté.
MAGISTRY Pierre, 2013, *Unsupervised Word Segmentation and Wordhood Assessment. Linguistics,* Paris Diderot.
MAO Gaga, 2019, « Study on Chinese Word Segmentation » in A*dvances in Higher Education,* Volume 3, Issue 3.
ORALOVA Gaisah, KUPERMAN Victor, 2021, « Effects of Spacing on Sentence Reading in Chinese » in *Front Psychol*, DOI: 10.3389/fpsyg.2021.765335 (consulté le 03/09/2022)

PEREA Manuel, WANG Xiaoyuan, 2017, « Do alternating-color words facilitate reading aloud text in Chinese? Evidence with developing and adult readers » in *Memory & Cognition*, no 45, p. 1160-1170.
PREVOT Laurent, MAGISTRY Pierre, HUANG Chu-Ren, 2015, « Un état des lieux du traitement automatique du mandarin » in *Faits de langues, 2015*.
SHEN Deli et al., 2012, « Eye movements of second language learners when reading spaced and unspaced Chinese text » in *Journal of Experimental Psychology Applied*, no 18(2), p. 192-202.
YANG-DROCOURT Zhitang, 2022, *L'écriture chinoise, au-delà du mythe idéographique*., Paris, Armand Colin.
YANG-DROCOURT Zhitang, 2007, *Parlons chinois*, Collection « Parlons », Paris, l'Harmattan.
YAO Panpan et al., 2023, « Explore the processing unit of L2 Chinese learners in on-line Chinese reading » in *Second Language Research*, DOI: https://doi.org/10.1177/0267658323120260 (consulté le 21/08/2024)
ZHOU Dongjie, ZHENG Zezhi, 2020, *Research on the Association Relationship Between Type-of-Words of Monosyllabic Word in the Modern Chinese*, Proceedings of the Third International Conference on Social Science, Public Health and Education (SSPHE 2019), p. 29-34.
ZHOU Wei, YE Wanwen, YAN Ming, 2020, « Alternating-color words facilitate reading and eye movements among second-language learners of Chinese » in *Applied Psycholinguistics*. 2020, no 41(3), p. 685-699.