Lignes brisées, recollées et démontées en linguistique informatique

Pierre Zweigenbaum

Lignes brisées, recollées et démontées en linguistique informatique

Pierre Zweigenbaum

Parole et écriture : une ligne continue continuellement brisée

« Le discours oral se déroule dans un flux temporel irrémédiablement linéaire. » (Vandendorpe, 25)

Ruptures de linéarité dans l'écriture concrète matérielle du matériau langagier

Écriture physique

Une rupture de linéarité s'est avérée nécessaire dans l'écriture pour faire tenir sur un support de taille bornée un texte trop long pour y être consigné d'une pièce. Déjà, comme le dit Vandendorpe (49), « sur la feuille de papyrus, qui était en usage depuis –3000, le scribe aligne les colonnes de texte en parallèle ». Ce papyrus est « enroulé sur lui-même en un volumen » (*ibid*.). Plus tard, le codex

> consiste en un ouvrage dont les feuilles pliées et reliées forment ce que nous appelons aujourd'hui un cahier ou un livre. Il est apparu quelques dizaines d'années avant notre ère dans la Rome classique à l'époque d'Horace […] Le passage [du rouleau au codex] ne sera vraiment effectué dans l'Empire romain qu'au IVème siècle (Vandendorpe, 50)

Enfin, « le Moyen Âge a aussi connu le rouleau, qui n'est plus le volumen mais le *rotulus* : il s'agit d'un long ruban de parchemin, mais écrit – et cette différence est fondamentale – non pas selon des colonnes perpendiculaires au ruban, mais dans le sens du ruban, sans colonne » (Larousse, 2020). Volumen, codex, *rotulus* brisent ainsi un texte en ce qu'il est convenu d'appeler des « lignes » (ou des colonnes selon les écritures), qui constituent des segments, des brisures de la ligne globale. Le flux de parole linéaire et continu évoqué plus haut est donc amené à apparaître comme ligne brisée lorsqu'il est écrit sur un support matériel. Notons que cette transformation apporte en pratique des avantages car « l'écrit nous permet d'échapper à la linéarité » (Vandendorpe, 1999, 40) : la forme tabulaire du codex facilite la lecture et nous permet de la mener à notre gré.

Lignes brisées et continues dans la gestion informatique du texte

Impression

L'image informatique d'un livre imprimé reproduit généralement ces segments qu'il faut alors savoir recoller de façon appropriée pour reconstituer la linéarité du texte. De nos jours, sa mise en page est calculée informatiquement selon les choix typographiques effectués (police, taille des caractères, taille de la page, etc.) : le nombre de mots approprié est disposé sur chaque ligne, des césures sont introduites si nécessaire ; le nombre de lignes approprié est inséré dans chaque page. Dans les publications de qualité, ce processus est contrôlé par un typographe professionnel qui effectue au besoin des ajustements. Le résultat est figé puis utilisé pour l'impression. Un document préparé pour l'impression, par exemple au format PDF (Portable Document Format), peut aussi être visualisé à l'écran. Dans ce cas, lignes et pages sont figées telles que préparées pour l'impression.

Affichage pour le lecteur

L'affichage d'un texte sur écran dans les logiciels modernes reprend également les principes de la page imprimée. Une différence avec le livre imprimé est que l'affichage à l'écran est susceptible d'être modifié dynamiquement. Le lecteur peut choisir par exemple d'augmenter la taille de la police de caractères, et le logiciel va ajuster en conséquence les points de repliement de la ligne du texte. Une autre différence est la possibilité souvent offerte de faire défiler le texte verticalement sur l'écran, reproduisant alors un *rotulus*. Page ou *rotulus*, l'affichage matériel du texte sur l'écran consiste ainsi en la matérialisation du texte sous la forme d'une ligne brisée dynamiquement ajustée.

Vue logique

Sur un plan informatique, les langages de programmation offrent au programmeur des structures de données pour enregistrer nombres, textes et autres données. Pour les textes, la structure de données standard est la notion de chaîne de caractères, ligne dont la taille n'est limitée que par celle de l'espace de stockage disponible. Dans les langages de programmation couramment utilisés en 2020 (par exemple, C ou Python), cet espace de stockage disponible est la mémoire centrale de l'ordinateur. Sur les ordinateurs personnels de 2020, cette mémoire est de l'ordre d'une dizaine de Giga-octets (1 Go ~ 1 milliard de caractères), ce qui représente environ mille livres de la taille de *À la recherche du temps perdu*. Bien que borné, cet espace est donc, contrairement à l'espace de l'écriture physique mentionné plus haut, suffisamment

grand pour contenir l'immense majorité des textes existants. La vue logique d'un texte fournie par les langages de programmation maintient ainsi virtuellement la ligne continue originale de la parole[48].

Stockage informatique « physique »

Sur un plan informatique plus profond, le système d'exploitation d'un ordinateur gère la mémoire disponible : comme l'explique Bloch (Chap. 4), il l'alloue par morceaux aux programmes qui lui font des demandes. Concrètement, il découpe la mémoire en blocs et tient à jour un annuaire des blocs libres et des blocs en cours d'utilisation. L'action conjuguée du système d'exploitation et du langage de programmation utilisé construit la vue logique, la ligne continue donnée aux programmes. Mais en sous-main, cette ligne est au besoin brisée par le système d'exploitation selon ses besoins de gestion de la mémoire de l'ordinateur.

Vue globale

Ces plans se combinent pour présenter, d'une part, au lecteur d'un texte informatisé une image virtuelle de ce texte, formatée en ligne brisée par imitation des pages d'un livre ou souvent d'un *rotulus*, et présenter, d'autre part, aux programmes qui le traitent une longue ligne reconstituée, tout en stockant souvent le texte brisé en morceaux dans la mémoire de l'ordinateur.

Les unités linguistiques : découper une ligne en une hiérarchie de segments imbriqués

Unités linguistiques

Un texte est écrit avec des caractères et constitue donc une séquence de caractères. En même temps, il est formé d'unités linguistiques. Chapitres, paragraphes, phrases, syntagmes, mots (voir notamment Swiggers) sont des éléments qui jouent un rôle à différents niveaux du fonctionnement de la langue, ou tout au moins de la modélisation qu'en fait la linguistique. Ces unités forment autant de segments de la ligne initiale et découpent cette ligne de façon imbriquée. Par exemple, un texte pourra être découpé en chapitres composés de paragraphes formés de phrases, chaque phrase étant elle-même une séquence de mots. Chacune de ces unités concerne une sous-séquence de la séquence de caractères formant le texte entier. On rencontre donc ici une autre forme de ligne brisée (figure 1).

48. Sur la notion de structure logique et structure physique d'un document, voir notamment André *et al.*.

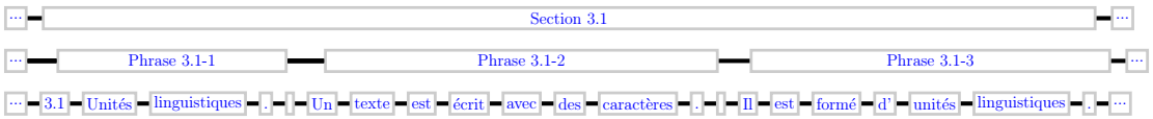


Figure 1. Lignes brisées : segmentation en unités linguistiques.

Segmenter un texte en unités linguistiques

L'analyse d'un texte, qu'elle soit humaine ou réalisée par des programmes de traitement automatique des langues[49], passe généralement par l'identification de ces unités et met alors en jeu un processus de segmentation du texte : la séquence de caractères, ligne continue initiale, est tronçonnée en une séquence de segments, chaque segment pouvant être à son tour découpé en segments plus petits. La figure 1 illustre cela au niveau des sections, phrases et mots d'un extrait de la section précédente.

Briser la linéarité du texte pour recouvrer ces unités, en d'autres termes, le segmenter, rencontre des difficultés d'autant plus grandes que les marques de ces segments sont diffuses ou que leur définition n'est pas ferme. C'est tout l'enjeu des processus de segmentation en mots, en phrases, en syntagmes dont même ceux qui semblent les plus simples, comme la délimitation des mots, sont confrontés à des questions théoriques et pratiques qui se répercutent dans les traitements en aval.

En l'absence d'une définition universelle, les programmes d'analyse automatique de textes adoptent des définitions opératoires qui résultent d'un compromis entre simplicité de mise en œuvre et utilité pour les traitements visés. La définition la plus répandue pour les langues où elle est pertinente consiste à prendre pour délimitation d'un mot les espaces et les ponctuations (figure 2a). Il faut éventuellement considérer que certaines ponctuations, comme le tiret en français, ne sont généralement pas des séparateurs de mots (figure 2b). Cette définition devient cependant rapidement trop étroite face aux diverses expressions polylexicales, aux phénomènes de figement et de non-compositionnalité très bien décrits par Mel'cuk. Des expressions comme *pomme de terre*, ou *en sous-main* et *au besoin* dans la (figure 2c), possèdent un degré de figement qui peut justifier leur segmentation en une seule unité. Les méthodes de traitement automatique des langues traitent ces cas à l'aide de grands dictionnaires d'expressions polylexicales tels que le DELAC constitué par Silberztein ou en estimant de façon statistique

49. Cori et Léon présentent une vue historique de l'emploi des termes proches *Traitement automatique des langues*, *Informatique linguistique* et *Linguistique informatique*.

le degré de figement d'une expression par l'observation de grands corpus avec des mesures comme celles de Church et Hanks.

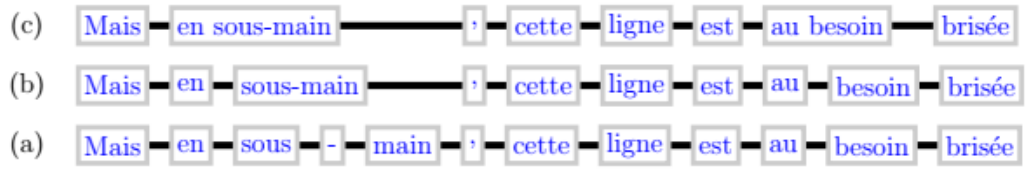


Figure 2. Lignes brisées : prise en compte d'expressions polylexicales dans la segmentation en unités lexicales.

Inversement, la présence d'erreurs orthographiques, la complexité de la morphologie de certaines langues, ou encore le besoin de limiter la taille du vocabulaire traité pour des raisons d'économie peuvent amener les concepteurs d'algorithmes d'analyse de textes à briser les énoncés en unités opératoires beaucoup plus courtes, comme des séquences infralexicales de quelques caractères.

Enfin, les langues dont l'écriture n'utilise pas d'espaces, comme le chinois (Wong *et al.*) ou le japonais, ou qui en utilisent entre toutes les syllabes, comme le vietnamien, font face de façon permanente à une question similaire à celle, introduite plus haut, des expressions polylexicales. La figure 3, qui reprend un exemple de Zhao *et al.*, illustre la segmentation d'une phrase chinoise en mots.

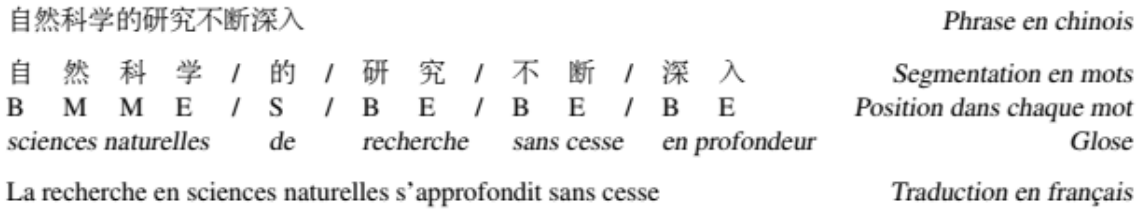


Figure 3. Exemple de segmentation en mots en chinois.

Les premières approches utilisaient un grand dictionnaire électronique et cherchaient la segmentation la plus en accord avec les mots présents dans le dictionnaire. La plupart des méthodes actuelles emploient des méthodes d'apprentissage automatique pour déterminer si dans le contexte où il est rencontré, un caractère est un mot en lui-même ou au contraire le début, le milieu ou la fin d'un mot à plusieurs caractères. La figure 3 symbolise ces positions par les lettres S (*single* = *seul*), B (*begin* = *début*), M (*middle* = *milieu*) et E (*end* = *fin*). Pour entraîner ces méthodes, les chercheurs leur fournissent de grands corpus qui ont été manuellement segmentés en mots.

Arbres et graphes se cachent derrière la linéarité de la parole

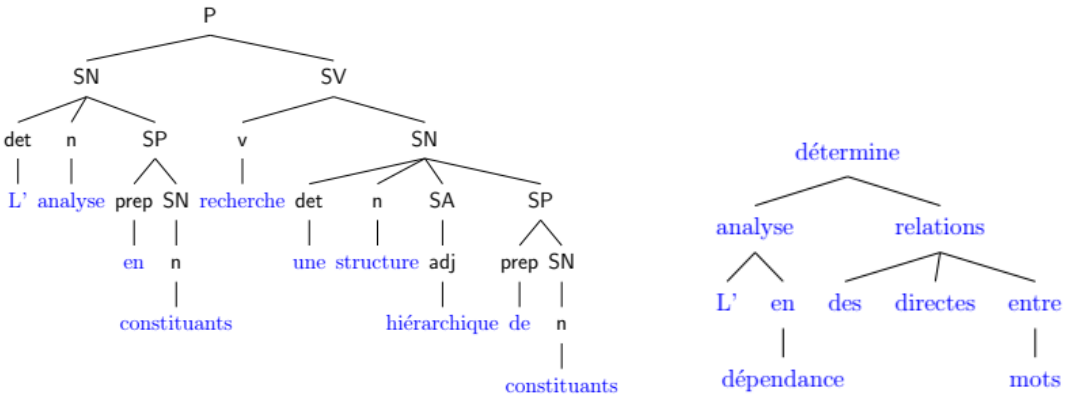
On peut voir une ligne brisée comme enchaînant des segments de ligne l'un derrière l'autre. Si au lieu d'ajouter un segment derrière le précédent on se permet d'en ajouter deux ou davantage, on obtient un branchement. Chaque fois que l'on répète cette opération, on crée un nouveau branchement. La structure résultante est un arbre : le point de départ du ou des premiers segments (dits *arêtes*) est la *racine* de l'arbre, et l'arrivée d'un segment sans suite est une *feuille*. La figure 4a montre un arbre dont les feuilles sont étiquetées par des mots et les branchements sont étiquetés par des catégories syntaxiques. Sa racine est notée P pour indiquer que l'arbre couvre globalement une phrase entière. Comme on le voit au passage, les linguistes, comme d'ailleurs les informaticiens, dessinent généralement les arbres en plaçant leur racine en haut et leurs feuilles en bas. On peut voir un arbre comme la connexion de segments de ligne : une ligne brisée que l'on recolle selon une structure arborescente, hiérarchique.

Un arbre est une sorte particulière de graphe : une structure composée d'*arêtes* qui relient des *sommets*. Alors qu'un graphe n'impose pas *a priori* de contrainte sur la disposition de ses arêtes, un arbre ne peut pas comporter de boucle, de circuit. Si dans un arbre on ajoute une arête entre deux sommets existants, cela crée une telle boucle, et la structure résultante n'est plus un arbre, mais un graphe plus général. On peut voir un graphe comme la connexion sans contrainte de segments de ligne : une ligne brisée que l'on recolle à son gré.

Structures non-linéaires

La linéarité de la parole et la quasi-linéarité de l'écriture sont des formes de transmission d'un contenu. Nous avons évoqué plus haut la notion d'unité de langue. Il faut y ajouter la notion de lien, de relations que les énoncés tissent entre ces unités aux niveaux syntaxique, sémantique et discursif, ainsi qu'entre ces unités et leurs référents dans une réalité extralinguistique. Les structures syntaxiques[50] sont des arbres (figure 4), voire des graphes plus généraux. Leur construction à partir d'un énoncé va donc transformer une ligne en arbre ou en graphe, en déterminant la structure appropriée et sa correspondance avec les éléments de l'énoncé. Inversement, la génération d'un énoncé à partir d'un contenu ou d'une structure syntaxique va devoir linéariser un graphe.

50. Voir par exemple le manuel d'Abeillé.



(a) Arbre de constituants (b) Arbre de dépendance

Figure 4. Arbres syntaxiques (P=phrase, SN=syntagme nominal, etc.).

Deux types d'arbres syntaxiques

L'analyse syntaxique part d'un énoncé, typiquement une phrase, et vise à déterminer la structure de cet énoncé.

L'analyse en constituants recherche une structure hiérarchique de constituants, ou syntagmes, imbriqués les uns dans les autres (figure 4a). La phrase est le syntagme englobant, constitué habituellement en français d'un syntagme nominal et d'un syntagme verbal, lui-même constitué notamment d'un verbe et de syntagmes nominaux ou prépositionnels, chaque syntagme se décomposant à son tour jusqu'à arriver aux mots. L'analyse en constituants est donc une autre sorte de segmentation, de niveau intermédiaire entre phrase et mot, qui porte sa propre hiérarchie. Elle représente chaque phrase par un arbre de constituants.

L'analyse en dépendance détermine quant à elle des relations directes entre mots (figure 4b). Chaque mot (régi) dépend d'un autre mot (recteur), sauf la racine de la phrase qui ne dépend d'aucun autre mot. Cela constitue un arbre de dépendance, partant de la racine de la phrase et dont les branches mènent aux dépendants, un dépendant pouvant être lui-même accompagné de ses propres dépendants, et ainsi de suite. Les relations de dépendance sont porteuses de relations grammaticales : sujet, objet, modifieur de nom, etc.

Arbre de constituants et arbre de dépendance sont d'ailleurs largement équivalents dans la mesure où chaque syntagme est organisé autour d'un mot de tête dont dépendent les autres

syntagmes qu'il inclut ; et réciproquement, chaque mot recteur constitue avec ses dépendants une unité qui correspond à un syntagme.

Constituants ou dépendances, la structure sous-jacente à une phrase est donc un arbre qui assemble des unités élémentaires en unités plus larges pour préciser, pour compléter l'apport de ces unités. Une structure où chaque mot serait attaché à celui qui précède construirait une séquence linéaire, une ligne. Le besoin de rattacher plus d'un dépendant à certains mots, par exemple à la fois un sujet et un objet à un verbe, ou à la fois un déterminant et un adjectif à un nom, ne peut s'exprimer dans une structure linéaire : il faut passer à une structure plus complexe, au minimum un arbre.

La phrase telle que nous l'entendons ou la lisons a été linéarisée en une séquence de mots pour la faire tenir dans le flux de parole. Mais de façon sous-jacente, c'est un arbre qui est exprimé et que nous reconstituons lorsque nous traitons l'énoncé reçu.

Autres types de relations : graphes

« Les phrases peuvent exprimer des relations qui vont au-delà des arbres que nous venons d'introduire. » Dans la phrase précédente, « phrases » est sujet de « peuvent », mais les propriétés lexicales du verbe « pouvoir » rendent « phrases » également sujet de « exprimer » : le sujet de « pouvoir » « contrôle » le sujet du verbe complément de « pouvoir ». De plus, dans la même phrase, le pronom relatif « qui » fait référence au nom « relations » et est sujet du verbe « vont » : par cette anaphore, au plan sémantique, « relations » est à la fois l'objet de « exprimer » et l'agent de « vont ».

« Les relations anaphoriques dépassent le cadre de la phrase. On les rencontre aussi entre deux phrases différentes ». Ainsi, « les » dans la phrase précédente fait référence à « Les relations anaphoriques » de la phrase d'avant ; et l'expression « la phrase précédente » fait référence à l'entièreté de la phrase en question. Par ailleurs, une phrase peut servir d'exemple à une autre : c'est le cas de la précédente (« Ainsi, … ») par rapport à celle qui la précède (« On les rencontre… »). Explicitation, argumentation, etc., de telles « relations discursives » structurent un texte, un discours, au-delà de la simple séquence linéaire de ses phrases, de la même façon que les relations syntaxiques structurent une phrase au-delà de la simple séquence linéaire de ses mots.

Ces relations s'ajoutent aux relations arborescentes vues au paragraphe précédent. La structure d'une phrase ou d'un texte est ainsi, au-delà d'un arbre ou d'une série d'arbres, un graphe qui relie mots ou unités plus larges selon les besoins de ce que l'on veut exprimer.

Délinéariser

Nous avons introduit plus haut la problématique de la segmentation d'un texte en unités linguistiques et pris l'exemple de la segmentation en mots. De la même façon, l'analyse d'une phrase cherche à identifier l'arbre ou le graphe qui définit la structure de cette phrase. C'est le cas notamment de l'analyse syntaxique et de ses versions computationnelles. Plusieurs décennies de recherches ont mis au point des algorithmes d'analyse syntaxique qui visent à construire automatiquement l'arbre structurant une phrase selon ses constituants ou ses dépendances. L'arbre de la figure 4b a ainsi été obtenu à travers l'interface en ligne de l'analyseur FRMG créé par de La Clergerie *et al*.

Le cheminement

Nous pouvons résumer cet essai en cheminant à partir d'un texte imprimé ou affiché sur un écran d'ordinateur. Ce texte découpe en pages et replie en « lignes », ou plus exactement en segments, la linéarité du flux de parole. En tant que lecteurs, nous recouvrons cette linéarité en parcourant le texte ligne après ligne, page après page. De façon sous-jacente, dans les traitements de l'ordinateur, le texte est également géré à un niveau logique comme une longue ligne. De façon encore plus profonde, au niveau du système de fichiers ou de la mémoire physique, cette longue ligne est en réalité fractionnée en segments qui peuvent être éparpillés ; mais la couche logique crée une abstraction qui masque ce fractionnement et fournit aux traitements une ligne d'une taille suffisante pour qu'y tienne n'importe quel texte courant. L'ordinateur réussit donc à gérer en parallèle la ligne brisée de la présentation à l'écran, la ligne différemment brisée de l'enregistrement dans sa mémoire physique, tout en préservant au plan logique la ligne continue du flux de parole.

Ce que nous avons besoin d'exprimer ne se satisfait cependant pas d'une simple structure linéaire. Nous brisons cette ligne encore différemment : en unités linguistiques, mots, syntagmes, phrases, chapitres et bien d'autres niveaux, souvent imbriqués les uns dans les autres. Ces relations d'imbrication forment des arbres. D'autres relations, comme l'anaphore, s'y ajoutent pour constituer des graphes. Ces relations complexes donnent au texte sa

cohérence, son unité, en identifiant notamment les personnages récurrents, les lieux, les situations, les assertions auxquelles on souhaite faire référence. La compréhension de textes a ainsi besoin de transformer la ligne de la parole en un ensemble d'unités unies par des relations complexes. Le traitement automatique des langues s'efforce de recouvrer automatiquement ces unités et ces relations pour simuler des aspects de cette compréhension.

Références

Abeillé A., *Les Nouvelles syntaxes. Grammaires d'unification et analyse du français*, Paris, Armand Colin, 1993.

André J, Furuta R. et Quint V., (éd.), *Structured Documents*, Cambridge, Cambridge University Press, USA, 1989.

Bloch L., *Splendeurs et servitudes des systèmes d'exploitation. Histoire, fonctionnement, enjeux*. Laurent Bloch, août 2020. En ligne, https://www.laurentbloch.org/Data/Livre-Systeme/, consulté le 7/10/2020.

Church K.W. et Hanks P., « Word Association Norms, Mutual Information, and Lexicography », *Computational Linguistics*, juin 1989, p. 76-83.

Cori M. et Léon J., « La constitution du TAL. Étude historique des dénominations et des concepts », *Traitement automatique des langues*, vol. 43, n°3, 2002, p. 21-55.

de La Clergerie E., Sagot B., Nicolas L et Guénot ML, « Évolutions d'un analyseur syntaxique TAG du français » in E. de La Clergerie et P. Paroubek, (ed.) *Journée de l'ATALA sur : Quels analyseurs syntaxiques pour le français ?*, Paris, France, 2009. ATALA. En ligne, http://hal.inria.fr/inria-00553260, consulté le 19/4/2020.

Encyclopérie Larousse, Entrée « volumen ». En ligne, https://www.larousse.fr/encyclopedie/divers/volumen/182698, consulté le 20/9/2020.

Mel'cuk I., « Clichés, an Understudied Subclass of Phrasemes », *Yearbook of Phraseology*, vol. 6, n°1, 2015, p. 55–86.

Silberztein M., « Le dictionnaire électronique des mots composés. », *Langue Française*, n°87, 1990, p. 71–83. En ligne, http://www.jstor.org/stable/41558558, consulté le 25/4/2020.

Swiggers P., « Le mot, unité d'intégration » *in* G. Serbat et S. Mellet, (dir.), *Études de linguistique générale et de linguistique latine : offertes en hommage à Guy Serbat*, Société pour l'information grammaticale, Paris, 1987, p. 57–64.

Vandendorpe C., *Du papyrus à l'hypertexte. Essai sur les mutations du texte et de la lecture*, Paris, La Découverte, 1999.

Kam-Fai Wong, Wenjie Li, Ruifeng Xu et Zheng-sheng Zhang, *Introduction to Chinese Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool, 2009.

Zhao H, Cai D, Huang C et Kit C, « Chinese Word Segmentation: Another Decade Review (2007-2017)« , *CoRR*, abs/1901.06079, 2019. En ligne, http://arxiv.org/abs/1901.06079, consulté le 6/5/2020.