Indices phonologiques des sinogrammes : de l'étude de l'acquisition à la modélisation pour l'apprentissage

Pierre Magistry[a] — Murielle Fabre [b,c] — Yoann Goudin[d]

[a] *LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay*
[b] *Collège de France - Unité de Neuro-imagerie Cognitive (INSERM U562)*
[c] *INALCO - CRLAO (CNRS-UMR 8563)*
[d] *INALCO - CERLOM (EA 4124)*
magistry@limsi.fr, muriellefabre@hotmail.com, yoanngoudin@yahoo.fr

RÉSUMÉ. *L'apprentissage d'une langue comme celui du mandarin présente un défi dont la difficulté principale consiste à saisir les correspondances entre les différents composants de la structure graphique du sinogramme et sa phonologie. En dépassant la stratégie didactique des listes de vocabulaire constituées sur des critères de fréquence, notre modèle veut présenter à l'apprenant les indices phonologiques et leur consistance au sein du système graphique dans sa globalité. À la frontière entre les disciplines, notre approche en TAL intègre dans le modèle présenté des propositions en didactique des langues ainsi que des résultats en psycholinguistique et neuro-imagerie.*

ABSTRACT. *Learning a language such as Mandarin Chinese includes specific challenges. A crucial point consists in grasping the right Orthographic-to-Phonology Correspondence (OPC) between the different graphical units in the sinogram and sound. Going beyond traditional vocabulary lists based on a lexical frequency strategy, we propose a computationnal model that enables to introduce the learner into the rules of the graphic system as a whole, its phonological cues and their reliability. At the crossroad between different disciplines, our NLP approach integrates the research results from Language teaching, Psycholinguistics and Neuro-imaging.*

MOTS-CLÉS : *acquisition, lecture, correspondances graphophonologiques, granularité, indices phonologiques, sinogrammes, mandarin.*

KEYWORDS: *reading acquisition, OPC, granularity, phonological cues, sinograms, Mandarin.*

1. Introduction

De par les défis spécifiques que l'écriture chinoise pose pour la numérisation des matériaux linguistiques, l'enseignement des langues qu'elle transcrit n'a pu commencer à bénéficier des résultats du traitement automatique des langues qu'assez récemment. Dans les dernières décennies, de plus en plus de travaux en TAL traitent et équipent le mandarin, comme en témoigne la multiplication des outils et ressources aujourd'hui disponibles pour cette langue (Prevot *et al.*, 2015). Les langues qui partagent avec elle l'usage du même système d'écriture et une part importante de leur lexique ne sont pas en reste : le japonais, le coréen, le taïwanais, langues ci-après désignées *langues sinogrammiques*[1]. Parmi les ressources disponibles, nombreuses sont celles diffusées librement. Ceci nous permet, en nous inscrivant dans la mouvance « *Linguistic Linked Open Data* », de les fusionner en une ressource liée que nous mettons à disposition de la communauté.

La particularité graphique des langues sinogrammiques a conduit à mettre au centre de l'enseignement-apprentissage des pratiques essentiellement grapholexicales. Toutefois, de récents travaux en sciences cognitives explorent la question de la *granularité* des systèmes graphiques, en étudiant les *correspondances graphophonologiques* de l'écriture chinoise et leurs indices phonologiques.

Inspirés par ces travaux, nous proposons ici de pallier l'absence d'une approche graphophonologique par la mise au point d'une modélisation devant permettre, à terme, de construire des ressources pédagogiques et d'être intégrée à des plates-formes de suivi de parcours d'apprentissage. D'un point de vue plus strictement didactique, nous cherchons à compléter les approches existantes, limitées par le manque d'un outillage formel, en définissant des *indices phonologiques*, et en précisant leur *granularité* sur les plans graphiques et phonologiques. Par ailleurs, contrairement à l'approche didactique dominante, dite *concentrée* (Bellassen, 2010), qui se fonde sur des seuils de vocabulaire et des listes institutionnelles basés sur la fréquence des caractères, notre modèle permet d'équiper une approche alternative. Son enjeu didactique est double : (1) présenter à l'apprenant l'économie interne du système graphique dans sa globalité dans une langue donnée, (2) exposer l'apprenant à la variation en synchronie à travers les différentes lectures de sinogrammes en langues affines, sans pour autant figer la programmation de l'apprentissage des sinogrammes.

La dimension multilingue au cœur de notre démarche n'est pas seulement le moyen de nous affranchir des pratiques nationales, comme les variantes graphiques des polices et d'encodages, mais nous prenons en compte un nouveau profil d'apprenants plurilingues. En effet, de plus en plus d'apprenants envisagent l'apprentissage consécutif voire simultané d'au moins deux de ces langues. Cette réalité confère ainsi aux sinogrammes et à leur empreinte phonétique un *potentiel d'intercompréhension* au fon-

1. Nous préférons l'usage du terme de *sinogramme* (Lyssenko et Anastasia, 1986) à celui de « caractères chinois ». Il permet d'aborder les relations *colingues* (Balibar, 1993) qui concernent aussi bien des langues – parfois typologiquement très distantes – que des variantes sinitiques telles que le taïwanais, le cantonais, etc.

dement de notre approche comme de la modélisation que nous proposons. Cependant, pour respecter les contraintes de longueur de cet article, nous illustrerons notre propos en utilisant essentiellement le mandarin transcrit au moyen du système de transcription officiel de la République populaire de Chine, le *Hànyǔ Pīnyīn*.

Dans cet article, nous nous attacherons tout d'abord à fournir des éléments de contexte sur le système graphique chinois pour ensuite présenter les enjeux didactiques de l'enseignement-apprentissage des sinogrammes. Dans la section 3, nous croiserons ces éléments avec les avancées récentes en sciences cognitives qui motivent les choix en terme de *granularité* pour notre modélisation. La section 4 présentera ensuite les ressources compilées pour alimenter le modèle. Enfin, la dernière section 5 présente notre modèle probabiliste qui rend compte de ce système de correspondances graphophonologiques en le faisant émerger des données.

2. Le système graphique chinois

Comme l'ont montré les recherches en grammatologie (Boltz, 1994) la centralité des *indices phonologiques* est constitutive du système graphique en diachronie. Nous rappelons ici les trois stades de développement proposés par Boltz. Ses travaux peuvent facilement être étendus à l'état actuel du système graphique, en considérant les pratiques de néologie. L'empreinte de ce phonétisme a depuis été observée en psychologie expérimentale, dans des travaux qui seront exposés en 3.1. Au-delà de ces éléments historiques, il s'agira de revenir ensuite sur les principes fondamentaux qui régissent l'économie du sinogramme et tout particulièrement sur la structure de la sino-syllabe ainsi que la composition graphique des sinogrammes.

2.1. *Un principe quasi exclusif de développement : le phonétisme*

Nous reprenons ici la thèse de Boltz (1994) qui distingue trois stades d'évolution de l'écriture et que nous illustrons dans la figure 1. Dans le prolongement d'un premier stade dit *zodiographique* – par opposition à pictographique car il s'agit déjà d'une représentation graphique qui ne permet plus la reconnaissance du référent – les signes sont arrivés à une seconde étape historique dite de la *multivalence*. À une certaine période de l'évolution du système graphique, certains signes sont devenus ambigus en raison du développement d'une polysémie. Ainsi, un *zodiogramme* donné pouvait être utilisé pour désigner son référent ou un signifié qui était homophonique comme illustré en A dans la figure 1. D'autres zodiogrammes, quant à eux, pouvaient être polyphoniques, un même signe pouvant être lu de deux façons différentes mais renvoyant à des concepts proches comme indiqué en B dans la figure 1. Enfin, dans un troisième stade dit *déterminatif*, les scripteurs ont finalement désambiguïsé les signes polysémiques et polyphoniques en associant à leurs formes initiales des composants dont la fonction était discriminante et permettait ainsi de distinguer des homophones devenant par ce procédé hétérographes et/ou préciser la lecture des sinogrammes dans la langue de ces scribes il y a 2 500 ans. Nous reprenons le terme *phonophore* de Boltz (1994)

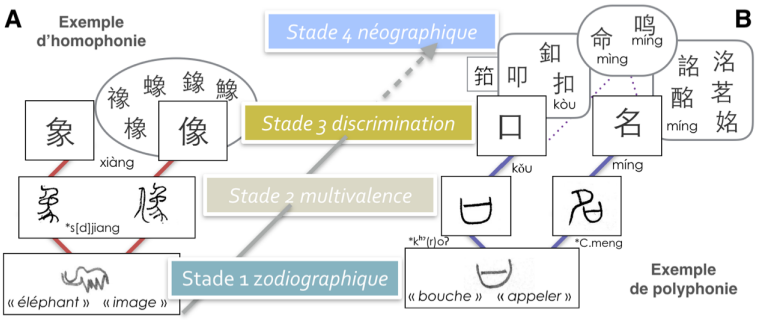

Figure 1. *Les différents stades de l'évolution de l'écriture chinoise d'après Boltz (1994, p. 69) avec la reconstruction en chinois archaïque (Ve AEC) précédée de '*' (Baxter et Sagart, 2014)*

pour désigner les composants dont la fonction renseigne sur la lecture du sinogramme considéré.

À ce troisième stade – et dernier dans l'exposé historique de W. Boltz – il faut ajouter un quatrième stade ou processus relatif aux besoins de la néologie chinoise depuis que le système s'est stabilisé, normalisé et institutionnalisé. Ce procédé de création « néographique » quasi exclusif consiste en la combinaison d'un sinogramme existant permettant de noter une syllabe homophone – ou quasi homophone – dont il n'existe pas de forme écrite connue du scripteur auquel est associé un composant discriminant.

Ainsi, au-delà de la linguistique historique, le constat principal de l'analyse du corpus des sinogrammes attestés montre qu'une écrasante majorité de ces derniers comportent un *indice phonologique*. En fonction de l'annotation, ce cas couvre entre 90 % et 97 % des dizaines de milliers de sinogrammes attestés (DeFrancis, 1989, p. 99).

Par ailleurs, ces observations permettent d'extraire des séries phonétiques dans lesquelles un même phonophore peut être utilisé en association avec différents composants discriminants, soulignant la centralité de ces indices phonétiques. Ainsi avec le *phonophore* 工 *gōng* « travail », il est possible de former une « série phonétique » – non exhaustive ici – comprenant 功 *gōng* « réussite », 攻 *gōng* « ennemi », 貢 *gòng* « tribut », 空 *kōng* « vide », 缸 *gāng* « jarre », 紅 *hóng* « rouge », 江 *jiāng* « fleuve », etc.

2.2. *L'économie du système sinographique*

L'économie du système sinographique repose sur deux éléments fondamentaux : le partage d'une structure syllabique à travers les différentes langues (tableau 1) et une structure graphique régissant l'agencement des différents composants du sinogramme.

INITIALE (*attaque*)	FINALE (*rime*)			
consonne		rime		
		ton		
	Médiale (*glide*)	Tonale (*nucléus*)	Terminale (*coda*)	
	j w y 0	voyelle	-m -n -ng -j -w 0	-p -t -k

Tableau 1. *Description de la sino-syllabe*

Les langues sinogrammiques présentent différentes solutions graphiques pour exprimer cette structure syllabique. Ainsi pour un sinogramme donné, indépendamment de sa réalisation graphique par exemple 寺 le « temple », il sera lu dans les langues sinitiques *sì* en mandarin, *sī* en taïwanais, en sino-coréen 사 *sa*, en sino-vietnamien *tự* [2] et en sino-japonais ジ *ji* [3].

L'autre dimension indéfectible de la précédente concerne l'analyse graphique des sinogrammes afin de dégager les principes fondamentaux qui régissent l'économie du système. Nous en traiterons ici deux aspects : les composants dont les sinogrammes sont constitués et leurs différentes fonctions (1) ainsi que le mode d'organisation spatial de ces composants dans un nombre fini de structures (2).

L'ensemble des sinogrammes est construit en combinant quelques centaines de composants, qui peuvent avoir différentes fonctions au sein d'un sinogramme donné.

Nous distinguons les quatre fonctions suivantes, exemplifiées avec le composant 寸 *cùn* signifiant le « pouce » :

– *phonophorique* (ou composant phonétique), comme dans 村 *cūn* le « village », où il est placé à droite avec le composant discriminant 木 dit du « bois » à gauche ;

– *discriminante* – alias « sémantique » ou « clé » – comme dans le sinogramme 寺 *sì* le « temple » où il est placé en position inférieure ;

– *graphique*, quand un composant est utilisé comme sous-composant, par exemple dans 時 *shí* « le temps » dans lequel 寸 n'est que le sous-composant du *phonophore* 寺 ;

– *sinogrammique* – par opposition aux trois autres fonctions *graphiques* – quand il s'agit d'un sinogramme indépendant tracé au centre d'un carré virtuel et attesté dans le lexique, par exemple 寸 *cùn* le « pouce ».

2. Quand bien même cette réalisation paraît très éloignée du chinois, on peut observer des correspondances systématiques entre les lectures sino-vietnamiennes et celles des autres langues sinogrammiques.

3. Tandis que le même sinogramme servira à noter le terme japonais pour dire « temple » てら *tera* : il ne s'agit plus là d'un emprunt à la langue chinoise.

Ces composants s'agencent entre eux au sein d'un carré imaginaire au centre duquel tout sinogramme est tracé. Ces types d'agencement de la structure interne du sinogramme sont au nombre d'une dizaine. Nous adoptons les douze configurations utilisées pour l'IDS – *Ideographic Description Sequence*[4] du consortium Unicode – décrit *infra* 4.1 : ⿰ ⿱ ⿲ ⿳ ⿴ ⿵ ⿶ ⿷ ⿸ ⿹ ⿺ ⿻ . Dans nos exemples supra, 村 *cūn* « village » et 時 *shí* « temps » relèvent de la structure ⿰ « gauche-droite », tandis que 寺 *sì* « temple » relève, quant à lui, de la structure ⿱ « supérieure-inférieure ».

2.3. *L'enseignement-apprentissage des sinogrammes*

Les pratiques et discours d'enseignement-apprentissage sont fondamentaux dans l'appréhension du système sinographique, et ce indépendamment du processus d'acquisition tel qu'étudié par les sciences cognitives. Il convient donc de fournir une brève présentation de l'enseignement-apprentissage des sinogrammes au sein des sociétés sinographiques puis en tant que langues sinogrammiques secondes.

2.3.1. *Langue(s) sinogrammique(s) langue(s) première(s)*

On remarquera tout d'abord que le sinogramme est le focus principal des cultures scolaires des pays en question et que l'enseignement L2 en porte une empreinte indélébile.

Discours et pratiques nationaux : reproduction et seuils

Dans les sociétés contemporaines concernées, l'apprentissage de la lecture et de l'écriture est programmé sur toute la durée de la scolarité obligatoire. À l'époque classique, cette programmation était implicite à travers un corpus de textes fondamentaux (Allanic, 2017). À l'époque moderne, sont apparues les institutions scolaires nationales qui ont chacune assigné des objectifs correspondant aux différents cycles de scolarité constituant autant de seuils de sinogrammes sélectionnés sur un critère de fréquence au sein du lexique de ces langues sinogrammiques légitimes. Le contenu et le nombre de sinogrammes pour ces seuils peuvent varier du simple au double depuis les 1 500 sinogrammes théoriques en Corée du Sud aux plus de 3 000 aussi bien en République populaire de Chine qu'en République de Chine (Taïwan), en passant par les 2 136 au Japon. En dépit de ces différences de valeurs, les pratiques pédagogiques sont fondées avant tout sur la mémorisation au moyen d'imitation d'un modèle – le livre ou le maître – de copie et de par cœur malgré des propositions raisonnées proposées par les didacticiens. La programmation et la division en seuils introduisent de manière précoce des sinogrammes dont la fonction dans d'autres sinogrammes sera avant tout discriminante. Cela s'explique surtout par le fait que cette connaissance est nécessaire pour trouver un sinogramme inconnu dans un dictionnaire traditionnel. Autrefois prérequis pour l'étude de la lecture, les usages du numérique sont en train de reléguer

4. Le choix malheureux du terme « *Ideographic* » est reconnu en tant que tel par l'Unicode mais y est conservé pour des raisons historiques.

ce procédé au rang de pratique spécialisée ou de curiosité historique. En dépit de ce biais en faveur des composants discriminants, on observe cependant qu'aujourd'hui en Chine populaire à la fin du cycle primaire, soit six années d'études, sur les 2 500 sinogrammes au programme, 72 % d'entre eux sont des *phonogrammes* 形聲字 *xíngshēngzì* dont les *indices phonologiques* offrent aux jeunes lecteurs des indices visuels distincts et fiables pour un accès au son et parfois au sens (Shu *et al.*, 2003).

2.3.2. *Langue(s) sinogrammique(s) langue(s) seconde(s)*

À destination d'un public allophone, outre la reproduction des discours et pratiques des enseignants natifs, il s'agit d'insister sur une spécificité de la didactique – en nous restreignant ici au seul mandarin langue étrangère en France – qui focalise essentiellement son objet sur la place, prépondérante, des sinogrammes dans l'enseignement-apprentissage (Bellassen, 2010).

Le paradigme didactique : l'approche concentrée

Il s'agit, aujourd'hui de l'approche didactique dominante, présentée par Joël Bellassen depuis la fin des années 1980 et qui demeure inchangée en 2017. Elle privilégie un traitement du lexique qui se veut très sophistiqué. En effet, ce dernier est programmé en fonction d'un double critère : sa pertinence communicative, pondérée par la haute combinatoire dans le lexique contemporain des sinogrammes qui constituent le terme en question. L'exemple canonique est le terme 可口可乐 *kěkǒukělè* « Coca-Cola », qui est important d'une part en tant que mot, du fait de sa pertinence communicative et, d'autre part – surtout – du fait de la très haute fréquence de chacun des trois sinogrammes dont ce terme est constitué. À l'inverse, un autre emprunt 咖啡 *kāfēi* « café », ne sera pas vu à l'écrit au motif de la rareté des sinogrammes qui constituent ce terme. Cette approche insiste tout particulièrement sur la notion de seuils de sinogrammes qui, de cadres pour borner l'apprentissage, sont devenus les listes officielles des programmes scolaires. Elles sont souvent interprétées et appliquées de façon très limitative à notre sens. L'adoption trop stricte de ces listes tend à repousser le recours à des documents authentiques aux niveaux les plus avancés.

L'approche globale

Sur ce constat, nous visons à construire une proposition alternative, que l'on nomme *approche globale*. Un de nos objectifs est de permettre l'accès aux documents authentiques à un stade plus précoce de l'apprentissage (Goudin, à paraître) ; (Goudin et Le, 2016). Pour cela, nous changeons le focus du sinogramme aux composants en rendant moins intimidante l'exposition de l'apprenant à des sinogrammes « hors programme ». Notre objectif global dépasse largement le cadre de cet article, qui se limite ici au travail nécessaire sur les phonophores. Ainsi, dans le domaine des savoirs sinographiques, l'approche globale ne conteste pas la productivité lexicale des sinogrammes tels que 可 *kě* et 口 *kǒu* dans « Coca (Cola) » pour composer d'autres entrées du lexique. En revanche, il est incontestable que les sinogrammes qui constituent « café » 咖啡 *kāfēi* sont également des candidats à un apprentissage précoce dès

lors qu'un changement de focus vers le composant est opéré. En conséquence, l'apprentissage des sinogrammes ne se limite pas au seul lexique d'apprentissage, mais au système graphique dans sa globalité dont tire le nom de cette approche. Ainsi pour l'exemple « café » 咖啡 *kāfēi* :

– le principe phonétique est évident : *kāfēi* est un emprunt transparent pour l'apprenant ;

– *leur structure gauche-droite* ⿰ *est la plus répandue* et donc rapidement familière des apprenants ;

– le couple structure gauche-droite et la distribution de la fonction des *composants discriminants* à gauche – ici 口 *kǒu* « bouche » – associée aux *phonophores* à droite – par exemple 加 *jiā* ou 非 *fēi* – est la *distribution fonctionnelle* la plus représentée du système ;

– les *indices phonologiques* – ou *phonophores* – 加 *jiā* et 非 *fēi* sont des sinogrammes de fréquence relativement haute dans le lexique contemporain en mandarin, signifiant respectivement « ajouter, plus » et « ne... pas » ;

– 加 *jiā* et 非 *fēi* sont par ailleurs des *phonophores* assez productifs rentrant dans la composition de sinogrammes fréquents du lexique contemporain en mandarin.

Bien sûr, toutes ces informations n'ont pas vocation à faire l'objet d'une maîtrise systématique et encore moins immédiate. L'enjeu et l'objectif pour l'apprentissage sont de pouvoir accéder le plus rapidement possible et en complète autonomie aux textes authentiques ainsi qu'à l'appréhension globale des principes qui régissent l'économie sinographique à commencer par le plus fondamental d'entre eux – le phonétisme – ce que ne permettent ni les approches natives, ni l'approche concentrée pour un public allophone.

3. Lecture du sinogramme : correspondances graphophonologiques et leurs indices

La recherche en psycholinguistique développementale sur la lecture des langues alphabétiques a montré que tout apprenti lecteur doit procéder à une analyse consciente de la structure du langage parlé que l'on nomme généralement *conscience phonologique*. Cette compétence joue un rôle majeur dans l'acquisition de la lecture , car elle permet d'identifier les éléments phonologiques des unités linguistiques – syllabes, attaque ou rime[5] d'une syllabe, ou bien encore phonèmes – et de les manipuler intentionnellement pour ensuite pouvoir les associer aux unités du système graphique lors de l'acquisition de la lecture. Ainsi, cette capacité à segmenter la chaîne parlée en unités plus élémentaires que les unités lexicales, et les représentations explicites et stables de ces unités phonologiques, s'avèrent indispensables pour que l'apprenti lecteur soit en mesure de découvrir ce qu'on appelle *les correspondances graphophonologiques* : les

5. À la terminologie linguistique *attaque* et *rime* correspondent en français les termes *initiale* et *finale* en phonologie traditionnelle chinoise (cf. tableau 1).

règles associant les propriétés graphémiques des mots – lettres et graphèmes – à leurs formes phonologiques (Windfuhr et Snowling, 2001).

La corrélation entre unités graphiques et phonologiques peut varier d'une culture graphique à une autre. L'apprenti lecteur de sinogrammes doit comprendre le principe régissant les correspondances graphophonologiques du système sinographique, où le calcul de la prononciation des sinogrammes n'est pas un assemblage son après son comme dans les écritures alphabétiques, mais plutôt un accès direct, à partir de ses différentes unités graphiques, à la phonologie de la syllabe et à ses sous-parties. Il instaure de cette façon une procédure de lecture qui consiste à convertir en sons les différentes unités graphiques qu'il trouve au sein de la structure interne du sinogramme (cf. section 2.2). Pour cela, il doit en identifier des éléments portant des *indices phonologiques* – les phonophores – et apprendre les règles de *correspondances graphophonologiques* du système sinographique, moins transparent que les écritures alphabétiques.

Pour illustrer ce point, nous reprenons l'exemple du *phonophore* 工 *gōng* « travail ». Lorsque l'on se trouve face à un sinogramme ayant le même *phonophore* qu'un autre, sa prononciation peut être strictement identique induisant une *homophonie parfaite* (1) – comme 功 et 攻 respectivement « réussite » et « ennemi » –, ou une *homophonie partielle* pouvant intervenir sur une ou plusieurs variables : *le ton*, comme dans 貢 *gòng* « tribut » (2), l'*initiale* comme dans 空 *kōng* « vide » (3a), la finale comme dans 缸 *gāng* « jarre », le ton et l'initiale dans 紅 *hóng* « rouge » (3b) (4), mais il peut également s'agir d'une syllabe toute autre (5) sans que ce composant ait perdu sa fonction d'indice phonologique renvoyant à une autre syllabe comme dans 江 *jiāng* « fleuve ».

Les enfants chinois apprennent à utiliser spontanément les indices phonologiques lorsqu'ils lisent de nouveaux caractères. Ces indices font, en effet, diminuer la charge mémorielle requise par l'apprentissage du large nombre de sinogrammes (Ho et Bryant, 1997a). On trouve, d'ailleurs, une corrélation significative entre *conscience phonologique* et l'utilisation des phonophores dans la lecture de sinogrammes et de pseudo-sinogrammes (Ho et Bryant, 1997b) ; (Ho *et al.*, 2000). Ces études permettent de confirmer que la sensibilité aux unités intra-syllabiques est essentielle pour l'apprentissage des règles de *correspondances régulières* entre la graphie chinoise et sa phonologie . Ces résultats suggèrent que les indices phonologiques qui sont à l'origine de la création d'une large majorité de sinogrammes (cf. section 2.3.1) sont bien utilisés par les jeunes apprentis lecteurs chinois, leur centralité pourrait donc constituer un élément fondamental du processus de décodage qu'implique la lecture chez tout apprenant.

3.1. *L'accès aux unités sub-syllabiques en lecture de sinogrammes*

Une des questions qui se pose dans notre approche est de saisir quelle granularité d'unités phonologiques correspond à l'utilisation des indices phonologiques chez les

apprentis lecteurs, en d'autres termes quelle est la taille des unités phonologiques que l'apprenti lecteur associe aux unités graphiques du système sinogrammique.

La lecture des indices phonologiques dans les langues sinogrammiques comme le mandarin présente plusieurs caractéristiques qui nous portent à nouveau au cœur de la structure interne de la sino-syllabe (tableau 1). Dans l'étude de la lecture, on identifie, en effet, plusieurs niveaux de conscience phonologique : la syllabe, l'attaque et la rime au sein de la syllabe, ainsi que le phonème. Or, chez les apprentis lecteurs de sinogrammes, ce sont la sensibilité aux unités syllabiques et aux initiales de syllabe qui ont été identifiées comme les meilleurs prédicteurs de la réussite ultérieure en lecture par rapport à la sensibilité aux phonèmes (Ho et Bryant, 1997a ; McBride-Chang et Ho, 2000). Plusieurs études (Ho *et al.*, 2004 ; Ho *et al.*, 2007), ont pu identifier dans la conscience phonologique du niveau de la rime et de l'attaque syllabiques une des toutes premières capacités nécessaires pour l'accès à la lecture.

Cela a permis de confirmer que la sensibilité aux unités intra-syllabiques est, en effet, essentielle pour l'apprentissage des règles de correspondances régulières entre la graphie chinoise et sa phonologie.

3.2. *Complexité visuelle et indices phonologiques*

La question de l'importance des capacités visuelles par rapport aux capacités phonologiques dans l'acquisition de la lecture se pose tout naturellement étant donné la nécessité d'un analyse visuelle fine de la structure interne des sinogrammes, pour y retrouver les indices phonologiques et autres composants graphiques. L'importance de ces deux capacités visuelles-orthographiques émerge plus particulièrement lorsqu'on analyse l'apprentissage ou la lecture experte des sinogrammes dits *irréguliers*, qui ne sont pas conformes aux correspondances graphophonologiques[6].

Les capacités visuelles prédisent le succès en lecture uniquement pour les enfants dans les premières années d'école et de 3 à 4 ans, alors que les mesures des capacités phonologiques (lecture de la transcription *pinyin*, habileté à discriminer les sinogrammes homophones et conscience phonologique de rime et d'attaque) prédisent les performances en lecture des enfants de 5 à 7 ans et en deuxième, troisième et cinquième années (Ho et Bryant, 1997a ; Ho et Bryant, 1997b ; Siok et Fletcher, 2001). Dans les classes supérieures, ce sont la conscience des tons des sinogrammes et la connaissance du *pinyin* qui sont corrélées aux capacités de lecture.

6. Il s'agit des sinogrammes dont la lecture n'est pas cohérente avec leur indice phonologique. Par exemple, le phonophore 寺 *sì* « temple » est considéré comme régulier dans 時 *shí* « temps » ou 詩 *shī* « poème », mais il n'en est pas de même dans 等 *děng* « classe, degré ».

3.3. *Granularité des correspondances graphophonologiques à travers les langues*

L'interrogation concernant la nature des unités impliquées dans les traitements nécessaires à la lecture des sinogrammes, au cœur de notre étude, n'a cessé d'être abordée par la littérature sur les causes sous-jacentes de la dyslexie, quelles que soient les cultures graphiques et les langues considérées. La recherche dans ce champ porte aujourd'hui plus particulièrement sur le degré d'incidence de la dyslexie développementale à travers les différentes cultures graphiques, non pas comme conséquence de la transparence des différentes orthographes, mais plutôt comme ayant des manifestations qui varieraient en fonction de la régularité orthographique de son écriture.

Pour illustrer cette hypothèse, nous citons ici le cas d'un lecteur bi-scriptural dyslexique manifestant des troubles de lecture en anglais mais pas en japonais (Wydell et Butterworth, 1999). L'origine de son déficit asymétrique a été reliée à la différence de granularité de l'orthographe entre les deux langues en question. Celle-ci serait donc un facteur central dans la compréhension des procédés de lecture.

En mandarin ou en cantonais, les dyslexiques ont tendance à prononcer aussi les sinogrammes irréguliers selon l'indice qu'offre leur phonophore. Ils se méprennent et prononcent 暗 *àn* « sombre » comme l'indiquerait son composant de droite 音 se prononçant *yīn* lorsqu'il est un sinogramme à part entière signifiant « son » (Fabre, 2009). À cela, s'ajoutent les erreurs dites d'analogie, où par exemple 跌 *diē* « tomber » est lu comme 铁 *tiě* « fer », ce qui consiste à lire un sinogramme selon la prononciation d'un autre ayant le même phonophore, alors qu'en réalité il devrait être prononcé sur la base de l'indice phonologique sub-syllabique partagé par les deux sinogrammes. L'analyse des déficits et des erreurs de lecture des dyslexiques fournit, ainsi, des indices sur la *granularité de représentation psychologique des sinogrammes*. Il s'agit là d'une méthode courante en psycholinguistique, permettant de mettre en lumière la granularité de l'unité de représentation psychologique sur laquelle se fonde le système de déchiffrage des correspondances graphophonologiques qui sous-tend à la capacité de lecture.

Ziegler et Goswami (2005) abordent ce point crucial en proposant que la capacité à stocker des matériaux phonologiques implique la nécessité pour le lecteur de trouver un dénominateur commun – la bonne granularité d'unités – entre, d'une part, le système de symboles que représente l'orthographe d'une langue et, d'autre, part sa phonologie, en vue de relier correctement les domaines graphique et phonologique. Ainsi, ils affirment qu'une théorie de la lecture se situant au niveau d'unités psycholinguistiques de traitement (*Psycholinguistic Grain-Size Theory*) serait plus adéquate pour établir une théorie universelle de la dyslexie se fondant uniquement sur l'interruption du développement phonologique, pouvant aussi expliquer le cas du dyslexique bi-scriptural reporté plus haut : les orthographes opaques obligeraient le système de lecture à développer des corrélations ayant diverses « *granulométries* » ou unités de traitement.

En résumant, tout comme les études présentées jusqu'ici, les erreurs des dyslexiques indiquent la contribution fondamentale des indices phonologiques et de la phonologie sub-syllabique au traitement à la fois normal ou déficitaire de la lecture

des sinogrammes. Cette double confirmation nous a amenés à sélectionner ces deux unités, justifiées d'un point de vue psycholinguistique, et de les placer au cœur de notre modèle.

3.4. *Valeur positionnelle des composants phonétiques : une contribution des sciences cognitives*

Parmi les propriétés les plus importantes des unités graphiques des sinogrammes – les composants dans leur ensemble, indépendamment de leurs différentes fonctions – on trouve celle relative à leur position dans sa structure interne (cf. 2.2). Les résultats expérimentaux rapportés dans cette section nous ont guidés dans la prise en compte dans notre modèle de cette propriété positionnelle. En effet, non seulement les enfants sont très tôt conscients de l'information positionnelle liée aux composants – ils savent la détecter lorsqu'on leur présente des suites de sinogrammes dès la deuxième année de scolarité (Shu et Anderson, 1999). Mais si l'on se penche sur la nature de la représentation mentale des unités graphiques internes aux sinogrammes, on observe (1) des effets sur les temps de latence corrélés au nombre de ces unités (Wang et Peng, 1997), et (2) différents effets d'amorçage positionnel et graphique. Ces résultats expérimentaux sont autant d'indices que la représentation des composants et leurs fonctions au sein des sinogrammes s'effectuent *via* un système de traitement visuel-orthographique, qui implique l'activation et l'accès à la représentation de chaque composant graphique contenu dans le sinogramme. Pour ce qui est des effets d'amorçage, on observe, en premier lieu, que la relation entre un composant graphique et son utilisation comme sinogramme indépendant a un effet facilitateur sur la lecture d'un sinogramme où il est composant graphique (Ding *et al.*, 2004). Comme représenté dans la figure ci-dessus,

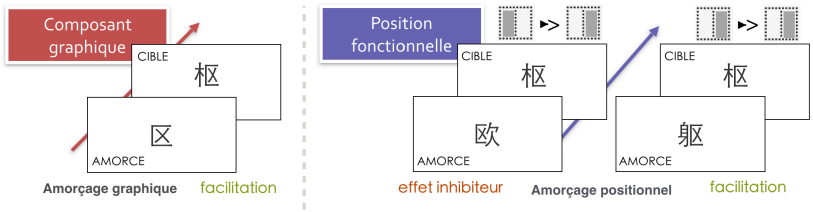

Figure 2. *Paradigme d'amorçage graphique et positionel (Ding et al, 2004)*

lorsque 区 *qū* « région » précède 枢 *shū* « pivot », ce dernier est lu plus rapidement, du fait d'un amorçage de la représentation du composant graphique de droite, qui pourtant n'est pas considéré en synchronie comme un *indice phonologique* proprement dit. Deuxièmement, on observe que lorsque deux sinogrammes partagent un même composant, la position fonctionnelle occupée par le composant dans la structure des deux sinogrammes, produit des effets d'amorçage différents. Comme illustré ci-dessus, lorsque 欧 *oū* précède 枢 *shū* « pivot », ce dernier est lu plus lentement, du fait d'un amorçage de la représentation du composant graphique à gauche qui entre en compétition avec l'activation de ce même composant à droite dans la cible 枢 *shū*. Au contraire,

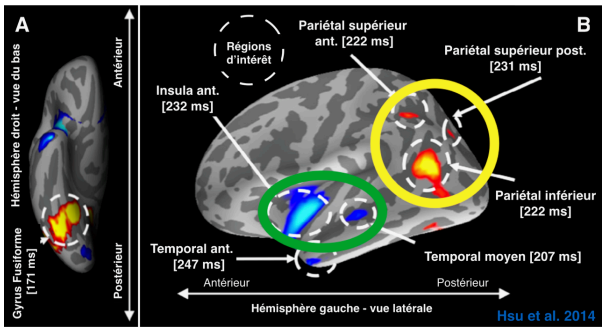

Figure 3. *Cartes des activations cérébrales moyennées (dSPM), reconstruction de sources en magnéto-encéphalographie (MEG) (Hsu et al., 2014)*

lorsque les composants se trouvent en même position dans le sinogramme qui précède 躯 *qū* « corps » et au sein de celui qui suit 枢 *shū*, ce dernier est lu plus rapidement du fait de l'amorçage de sa représentation et de sa préactivation en position droite.

3.5. *Cohérence et fréquence de l'indice phonologique du sinogramme*

Les données en neuro-imagerie viennent aussi alimenter nos réflexions et confirmer nos positions. Une étude récente s'est penchée sur deux caractéristiques centrales dans notre approche (Hsu *et al.*, 2014) : (1) la fréquence des indices phonologiques, définie comme le nombre de sinogrammes partageant le même indice, et (2) la cohérence phonologique au sein du système de ces indices, calculée comme le rapport du nombre de sinogrammes ayant la même prononciation et le même indice phonologique au nombre de sinogrammes présentant l'indice phonologique en question.

Les auteurs rapportent que l'effet de fréquence combinatoire des phonophores est observable dans une aire cérébrale distincte de celle liée à la mesure de cohérence phonologique (figure 3A cercle blanc en pointillé, le gyrus fusiforme droit) à 170 ms après la lecture silencieuse d'un sinogramme présenté à l'écran. Cette activation est interprétée comme le reflet de l'expertise perceptuelle dans le traitement de l'orthographe des sinogrammes. Il s'agit, en effet, de l'homologue droite de l'aire de la forme visuelle des mots, une aire cérébrale spécialisée dans le traitement des aspects visuo-graphiques des mots, et leur identification visuelle. Un deuxième résultat intéressant pour notre étude est rapporté dans une fenêtre temporelle successive (200 ms – 250 ms) (figure 3B cercle vert) où l'on observe une plus forte activation pendant la lecture de sinogrammes avec des indices phonologiques de fréquence plus basse par rapport à ceux de plus haute fréquence. À cela s'ajoute l'observation d'une activation cérébrale encore différente (cf. figure 3B cercle jaune) pendant la lecture de sinogrammes ayant des indices moins cohérents par rapport à ceux avec des indices plus cohérents à l'échelle du système graphique. Cette dernière aire cérébrale – le cortex pariétal inférieur gauche

– est connue pour être impliquée dans le traitement des correspondances graphophonologiques de la lecture à travers les langues. Les différentes localisations et fenêtres temporelles des activations cérébrales rapportées par les auteurs permettent de mettre en lumière une séparation fonctionnelle des différents traitements liée aux mesures de fréquence et de cohérence des indices phonologiques du mandarin.

Les résultats présentés dans cette section ont motivé un certain nombre de choix dans la conception de notre modèle, comme l'intérêt d'un point de vue cognitif de se concentrer sur le niveau sub-syllabique et de prendre en considération un paramètre indiquant la fiabilité pour le lecteur des indices phonologiques. De façon plus générale les résultats expérimentaux présentés suggèrent, que si les phonophores trouvent leur origine dans la genèse même du système graphique chinois, ils jouent un rôle central dans le déchiffrement qu'implique la lecture experte en langue sinogrammique. Il nous semble donc important de pouvoir capturer ce phénomène pour le présenter aux apprenants. C'est le rôle du modèle que nous allons maintenant détailler.

4. Ressources

Avant d'entrer dans les détails de notre modélisation, soulignons que celle-ci est grandement facilitée par l'abondance de ressources librement disponibles décrivant les langues qui nous intéressent ainsi que diverses informations sur la graphie des sinogrammes. Ces ressources sont d'origines diverses et diffusées dans des formats variés. Les efforts qui vont dans le sens des bases lexicales ouvertes et liées (Gurevych *et al.*, 2016) sont essentiels pour des travaux comme les nôtres. Pour cet article, nous bénéficions d'une version non finalisée d'un réseau de connaissances autour des sinogrammes et des langues sinogrammiques proposé dans (Magistry *et al.*, 2015) qui s'appuie sur les standards du Web sémantique (RDF et OWL). Nous citons ici les ressources desquelles sont originaires les données et leurs principales caractéristiques. On ne présentera cependant pas l'ensemble de la base, qui s'étend bien au-delà des objectifs de cet article, mais nous nous contenterons d'illustrer les informations qui nous sont utiles pour notre modèle à la section suivante.

4.1. *Descriptions IDS*

Le projet CHISE[7] (*Character Information Service Environment*), conduit au sein de l'AIST[8] japonais, visait à organiser les connaissances autour de l'écriture et à fournir des outils pour les manipuler informatiquement, notamment pour les besoins des recherches en humanités numériques. Il s'est étalé de 2002 à 2012 mais certains de ses sous-projets sont encore actifs. Nous utilisons l'ensemble des descriptions de la graphie d'un grand nombre de sinogrammes en IDS (*Ideographic Description Sequence*) fournies par CHISE. L'IDS est un formalisme de description de la structure interne

7. `http://chise.org`
8. National Institute of Advanced Industrial Science and Technology.

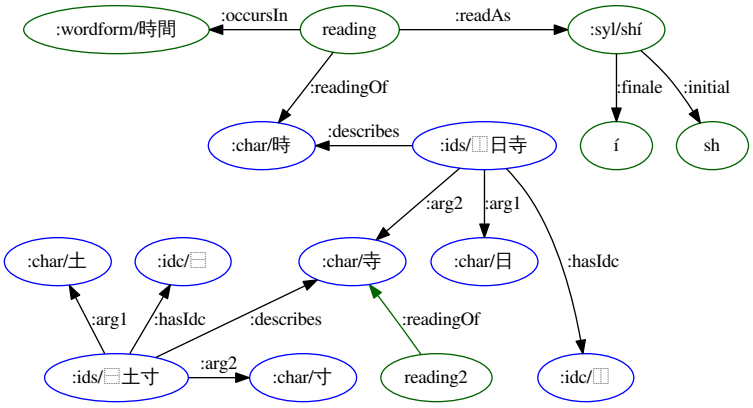

Figure 4. *Sous-graphe RDF construit à partir des exemples donnés à la section 2.2 et « liés » au reste de nos données : “ 寸 → 寸 ”, “ 寺 → ⿰土寸 ” et “ 時 → ⿰日寺 ”. La partie en bleu correspond aux données extraites des IDS de CHISE, la partie en vert provient des ressources lexicales.*

des sinogrammes, adopté par le consortium Unicode. Il est notamment utilisé pour désigner les caractères absents de l'Unicode ou pour discuter du statut des variantes « nationales » d'un sinogramme qui peuvent mériter d'avoir des codes distincts ou correspondre à un même code qui sera rendu légèrement différemment par des polices suivant les normes de différents pays. Dans le cadre du projet CHISE, un travail de description impressionnant a été accompli puisque pour les seuls caractères inclus dans l'Unicode, plus de 80 000 ont été décrits.

Une description en IDS commence par un des douze opérateurs de composition introduits en section 2.2, nommés IDC (pour « *Ideographic Description Characters* »). Ils permettent de combiner entre eux deux ou trois caractères existants ou d'autres IDS construites de la même manière (récursivement). On obtient ainsi un arbre de décomposition dans lequel la racine est un IDC et les feuilles sont des composants. Pour éviter le recours à des parenthèses, l'IDS utilise une notation polonaise inversée où l'opérateur est placé avant ses arguments. On analyse ces descriptions afin de les intégrer à notre base RDF. Elles sont ainsi directement stockées sous forme d'arbres interconnectés et liées au reste de nos ressources comme l'illustre la figure 4.

4.2. *Ressources lexicales*

Notre base compile des informations lexicales d'origines diverses[9]. Elles nous fournissent notamment des informations sur les lectures possibles des sinogrammes

9. Les données du dictionnaire du mandarin 《重編國語辭典修訂本》, consultable à l'adresse `http://dict.revised.moe.edu.tw/`) sont disponibles en libre téléchargement à

observés dans des expressions plus larges, aux emplois attestés dans les corpus de langues modernes. Dans les ressources d'origine, les expressions (d'un ou plusieurs sinogrammes) sont associées à leur prononciation (d'une ou plusieurs syllabes) dans une forme translittérée, le plus souvent une romanisation. Dans notre graphe de données RDF, on établit les correspondances au niveau syllabique tout en gardant un lien avec l'expression d'origine. On enrichit un peu plus le graphe en indiquant comment se décompose chaque syllabe. Pour les expériences présentées ici, on met de côté les informations sémantiques et morphosyntaxiques.

5. Modélisation

Les ressources que l'on vient de décrire nous permettent de modéliser les connaissances détaillées dans les sections précédentes.

On a vu que des modèles proches du nôtre ont été proposés dans le cadre d'expériences en sciences cognitives (Lee *et al.*, 2010 ; Hsu *et al.*, 2014), avec pour but de trouver des corrélations entre les indices calculés et les activations cérébrales. Cependant Lee *et al.* (2010) se situent au niveau du sinogramme et proposent un « indice de cohérence », tandis que Hsu *et al.* (2014) ne prennent en compte qu'une unique prononciation (la plus fréquente, en mandarin) pour un phonophore donné. Sans entrer dans les détails, la différence la plus marquée avec notre travail, qui justifie à elle seule la proposition d'un nouveau modèle, est qu'un score est attribué à chaque sinogramme, tandis que notre objectif est de cibler l'indice phonologique pour se doter d'un moyen d'estimer la fiabilité de chaque phonophore. Dans les sections précédentes, nous avons présenté nos motivations pour dépasser ces limitations. Dans cette section, nous exposons les détails de notre calcul. Un modèle formel et son implémentation informatique nous permettent d'envisager des applications plus variées qu'avec une simple liste de phonophores qui aurait été compilée manuellement par un ou plusieurs individus érudits. Le modèle présenté dans la suite de cette section est relativement simple. Toutefois, différentes options d'implémentation et de sélection des données utilisées pour des calculs vont produire différentes variantes du modèle qui peuvent ainsi correspondre à différents cas d'utilisation. On pourra ainsi se focaliser sur différents ensembles de langues, différents corpus de sinogrammes (programmes d'enseignement), ou même aller vers la modélisation des connaissances d'un apprenant spécifique si l'on suit sa progression dans le lexique en langue cible.

---

l'adresse `http://resources.publicense.moe.edu.tw/dict_reviseddict_download.html`. Celles pour le taïwanais sont consultables à l'adresse `http://twblg.dict.edu.tw/` et peuvent être obtenues sur demande (voir les instructions `http://twblg.dict.edu.tw/holodict_new/compile1_6_1.jsp`)

5.1. *Estimation de la fiabilité d'un indice phonologique*

Un composant c peut être considéré *phonophore* si (et seulement si), il apporte de l'information sur la façon de lire les sinogrammes dans lesquels il apparaît. En supposant connue une loi de probabilité P des différentes lectures, ce gain d'information est directement modélisable par une baisse de l'entropie (H) dans le cas où l'on sait que c est composant du sinogramme lu. Ce n'est ni plus ni moins que la définition du *gain d'information* en théorie de l'information (Shannon, 1948) :

$$IG(P,c) = H(P) - H(P|c)$$

Dans notre cas, $H(P)$ sera une constante. Comme on s'intéresse moins à la valeur exacte de $IG(P,c)$ qu'à la relation d'ordre qu'elle définit pour classer les graphèmes suivant leur fiabilité en tant qu'indice phonétique, il nous suffit de mesurer $H(P|c)$ pour chaque graphème c. On obtient une valeur continue, ce qui correspond bien à notre objectif de ne pas effectuer un classement binaire entre phonophores et non-phonophores. On veut pouvoir dire que certains phonophores sont plus fiables que d'autres pour inférer une prononciation. Une faible entropie conditionnelle correspondra donc à un indice fiable (un *phonophore*) tandis qu'une entropie élevée correspondra à une absence d'indice (un composant *discriminant*).

5.2. *Estimation de la probabilité des lectures*

Notre modélisation de la contribution phonologique d'un composant se base donc sur l'idée simple de recourir au gain d'information $IG(P,c)$. Mais pour pouvoir effectuer ce calcul, il nous faut encore préciser un certain nombre de détails et de choix d'implémentation. On doit notamment se doter d'une façon d'estimer P, et aussi préciser ce que l'on retient formellement comme « composants », candidats au statut de phonophore. Cela va nous amener à complexifier le modèle pour coller au mieux aux observations des sections précédentes.

5.2.1. *Aspects phonologiques*

Les réalisations phonologiques des lectures sont des *sino-syllabes*, mais comme on l'a vu section 2.2, celles-ci peuvent être décomposées de différentes manières. La structure interne de la *sino-syllabe* nous permet d'estimer de façon distincte la probabilité d'observer telle ou telle sous-partie (initiale, finale, attaque, rime, noyau, coda) et de définir la probabilité d'observer une syllabe comme une probabilité jointe ou comme plusieurs probabilités.
Cette approche limite l'effet de dispersion des observations et correspond mieux à notre objectif de capturer des phonophores qui sont susceptibles de n'apporter de l'information que sur une sous-partie de la syllabe comme la psycholinguistique nous l'apprend (voir section 3.1). On choisit donc d'effectuer deux calculs d'entropie distincts, un pour l'initiale, l'autre pour la finale et l'on en conserve la somme.

Si l'on note FP la fiabilité d'un phonophore, P_i la probabilité des initiales et P_f la probabilité des finales, on cherche donc :

$$FP(c) = H(P_i|c) + H(P_f|c)$$

5.2.2. *Aspects graphiques*

De la même manière que pour la syllabe, les graphies peuvent être décomposées de différentes façons. Les choix sur la façon de décomposer vont affecter l'ensemble des composants considérés comme phonophores potentiels, et par conséquent la façon dont les comptes sont effectués pour les estimations d'entropie conditionnelle. Comme présenté en section 4, l'IDS nous fournit une décomposition arborescente qui peut être effectuée récursivement. On peut donc considérer plusieurs « niveaux » de décomposition. La position du composant est aussi indiquée par l'IDC et la position de l'argument.

L'histoire du développement des sinogrammes (section 2) et les récents travaux en neuro-imagerie nous poussent à considérer essentiellement le premier niveau de décomposition (suivant le modèle de construction par ajout d'un composant discriminant à un sinogramme homophone, l'homophone devenant phonophore). Par ailleurs, on a vu que la position du composant avait son importance pour lui attribuer ou non le statut de phonophore. Pour capturer cette position, on combine le type d'IDC utilisé pour décrire la composition des sinogrammes et la position argumentale du composant considéré. Ainsi de 時 , dont la décomposition est décrite ⿰日寺 , on peut extraire deux composants : ⿰日 x et ⿰ x 寺 . Pour chacun des composants ainsi définis, on peut extraire de notre graphe RDF tous les sinogrammes comportant le composant avec une partie x quelconque.

5.2.3. *Intégration du multilinguisme*

Notre base de données et une partie de nos objectifs sont fortement multilingues. Il nous faut donc faire attention à prendre en compte le fait que pour chaque composant, nous avons accès à des lectures variées dans plusieurs langues. De plus, chacune des langues est décrite par des ressources dont les tailles peuvent varier grandement d'une langue à l'autre. Avec la formulation précédente, on court le risque de défavoriser les phonophores qui sont actifs dans plusieurs langues, car cette diversité des lectures aurait tendance à faire augmenter l'entropie.

La solution que l'on adopte est d'effectuer des calculs distincts pour chaque langue, et de les réunir en une moyenne par composant. On propose d'effectuer ces moyennes sur des scores normalisés par langue au moyen d'un z-score pour être moins sensible aux tailles relatives des ressources lexicales. On obtient ainsi la formulation (finale) de la fiabilité d'un phonophore c qui suit :

$$FP_{\mathcal{L}}(c) = \frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} z_l(H(P_{i,l}|c) + H(P_{f,l}|c)),$$

Où $\mathcal{L}$ est l'ensemble des langues considérées, $P_{i,l}$ (resp. $P_{f,l}$) sont les probabilités des initiales (resp. finales) dans une langue l donnée. Enfin z_l dénote une fonction de normalisation pour une langue l donnée, $z_l(x) = \frac{x-\mu_l}{\sigma_l}$ où μ_l est la moyenne de l'ensemble des valeurs obtenues pour tous les composants dans la langue l et σ_l la déviation standard du même ensemble de valeurs. On obtient ainsi une mesure empirique qui crée une relation d'ordre entre les composants et nous permet de classer ceux-ci du meilleur au moins bon phonophore. On peut alors enrichir notre graphe de données avec les résultats obtenus ou laisser les futurs utilisateurs effectuer eux-mêmes les calculs si des paramètres ou des données doivent être modifiés.

5.3. *Implémentation*

Les calculs présentés ci-dessus et les expériences pour l'évaluation du modèle décrites ci-dessous ont été implémentés en Scala en utilisant la bibliothèque Jena pour manipuler la base RDF. L'extrait de la base et le code source commenté pour reproduire les résultats présentés dans cet article sont disponibles à l'adresse `https://a-tsioh.github.io/Phonophores`

5.4. *Évaluations*

5.4.1. *Évaluation qualitative*

Il est assez difficile de proposer une évaluation quantitative rigoureuse pour le modèle proposé. C'est une des raisons pour lesquelles les motivations psycholinguistiques sont essentielles. Des observations quantitatives sont données plus bas mais avant cela, on peut aussi observer les valeurs obtenues pour quelques exemples utilisés dans les sections précédentes dans le tableau 2. Ici, nous utilisons donc un modèle estimé sur la base du lexique mandarin.

On observe que les valeurs relatives des composants correspondent assez bien à nos attentes. Un survol de la liste des composants enrichie de notre indice de fiabilité phonologique semble confirmer l'adéquation entre la sortie du modèle, nos motivations et nos objectifs.

5.4.2. *Évaluation quantitative*

Il n'existe pas de jeu de données annotées de référence qui correspondent à notre problématique. Cependant, lors de travaux antérieurs, nous avions annoté une longue liste de sinogrammes décomposés chacun en deux sous-composants en indiquant si le premier ou le second composant était phonophore[10]. On utilise ici ces annotations pour calculer la proportion de fois où un composant est phonophore par rapport au nombre

10. Cette annotation avait été effectuée en prenant en compte les lectures en mandarin et les décompositions données par Wikipédia. Nous leur préférons aujourd'hui les décompositions de l'IDS, plus standard et en plus grand nombre. Cependant les compositions haut-bas, gauche-

1	⿰ x 寺	1,12	時 shí 持 chí 峙 chí 峙 zhì 侍 shì 特 tè 恃 shì 詩 shī
2	⿰ x 寸	2,33	跱 zhì 待 dài 待 dāi 畤 zhì 尉 wèi 尉 yù 村 cūn 肘 zhòu 對 duì 忖 cǔn 討 tǎo 耐 nài 封 shù 紂 zhòu 封 fēng 射 yè 射 shè 射 yì 射 shí 吋 cùn 付 fu 付 fù 耐 zhòu
3	⿱ x 寸	1,68	奪 duó 專 zhuān 專 fū 寺 sì 尊 zūn 守 shòu 守 shǒu 導 dào 導 dǎo 辱 rù 辱 rǔ
4	⿰ x 加	0,31	刹 jiā 伽 qié 伽 jiā 跏 jiā 枷 jiā 咖 kā 珈 jiā
5	⿰ x 非	0,21	啡 fēi 琲 bèi 靡 fēi 鯡 fèi 誹 fěi 騑 fēi 徘 pái 裶 fēi 悱 fěi 排 pai 排 pǎi 排 pái 俳 pái 緋 fēi 腓 féi
6	⿰ x 音	− 0,75	黯 ān 愔 yīn 喑 yīn 揜 ǎn 腤 ān 暗 àn 韽 ān 諳 ān
7	⿰区 x	− 0,62	鸥 ōu 欧 ǒu 欧 ōu 殴 ōu 瓯 ōu
8	⿰ x 区	1,47	讴 ōu 岖 qū 妪 yù 呕 òu 呕 ǒu 呕 ōu 呕 xū 伛 yǔ 沤 òu 沤 ōu 眍 kōu 枢 shū 抠 kōu 驱 qū 躯 qū

Tableau 2. *Résultats obtenus par nos modèles sur les composants mentionnés au long de l'article (dans l'ordre où ils apparaissent dans notre texte)*

de fois où il apparaît dans un sinogramme. On peut alors comparer ces valeurs à nos scores de fiabilité et obtenir la figure 5. Dans celle-ci, chaque point représentant un composant est placé suivant la proportion obtenue manuellement (abscisse) et le score attribué par notre modèle (ordonnée).

Il est clair que ces deux scores sont construits de manières très différentes. On ne peut donc pas les comparer pour effectuer une évaluation comme il est coutume de le faire en TAL. Cependant ces deux scores décrivent deux aspects des mêmes objets qui ne sont pas sans rapport puisque les deux portent sur les phonophores. Il est ainsi satisfaisant d'observer une forte corrélation ($r = -0{,}64$) avec une très forte significativité statistique ($p << 0{,}001$) entre les données annotées manuellement et les scores obtenus automatiquement.

5.4.3. *Liste du ministère*

Le ministère de l'Éducation nationale publie au *Bulletin officiel* les listes de sinogrammes qui sont au programme de mandarin dans le secondaire. En utilisant notre modèle, on peut chercher à vérifier si la liste des sinogrammes considérés comme essentiels permet de saisir la logique du système graphique et cette relation entre la graphie et la phonologie. Pour cela, on compare un modèle construit sur la base d'un lexique du mandarin standard complet et un modèle construit en se limitant aux lectures des sinogrammes présents dans le programme de la LV1 (au nombre de 805). On présente les résultats à la figure 6. On voit que le modèle ne peut pas capturer les correspondances graphophonologiques en se basant uniquement sur les composants

droite et extérieur-intérieur sont communes et couvrent déjà une très large majorité des sinogrammes.

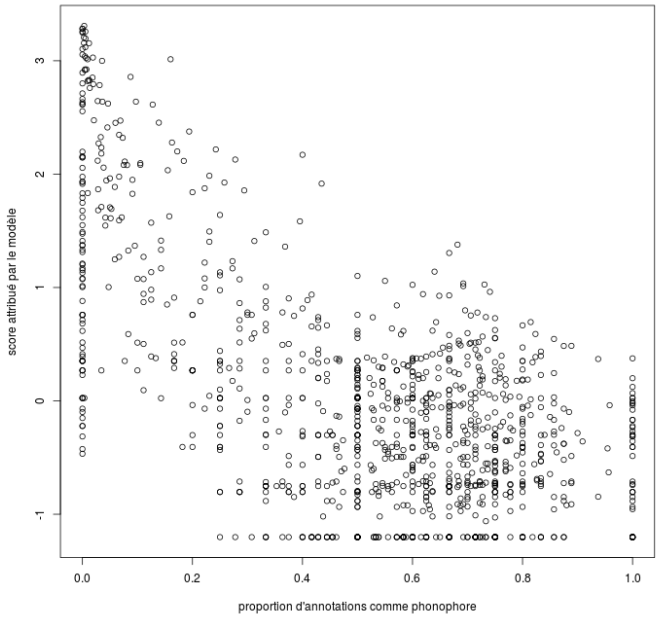

Figure 5. *Corrélation entre la proportion de sinogrammes pour lesquels un composant est phonophore et le score donné par notre modèle. ($r = -0{,}64$, $p \ll 0{,}001$*

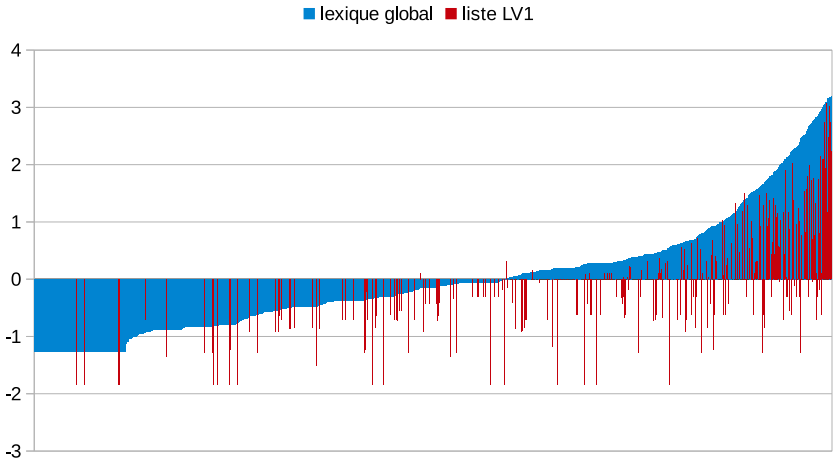

Figure 6. *Composants triés par fiabilité de phonophores, calculée sur le lexique dans sa globalité (en bleu) ou limitée par la liste du ministère pour la LV1. Le tri suit l'ordre du modèle global.*

présents dans les 805 sinogrammes de la liste de LV1[11]. On observe que pour ce qui est couvert par cette liste (en rouge), la zone la moins dense se situe dans les valeurs très négatives à gauche, c'est-à-dire que les phonophores les plus fiables sont absents du programme d'enseignement. À l'inverse, la zone de droite, positive pour le modèle global est bien mieux couverte mais concerne les composants qui ne sont pas phonophores. On constate aussi que le modèle LV1 attribue des valeurs très négatives à des composants qui n'ont pas de contribution phonologique, dénotant une régularité dans le programme de LV1 qui ne se vérifie pas sur l'ensemble du lexique. Les résultats que nous obtenons indiquent que la liste des sinogrammes de LV1 n'est pas représentative du fonctionnement global du système graphique pour ce qui est des correspondances graphophonologiques. Cela montre que l'enseignement-apprentissage qui s'appuie sur de telles listes ne peut pas s'en contenter et doit nécessairement aller au-delà afin de permettre à l'apprenant de devenir autonome dans l'acquisition du lexique et sa spécialisation.

6. Conclusion

Les résultats présentés dans cet article sont un exemple de dialogue interdisciplinaire entre didactique, sciences cognitives et traitement automatique des langues. Nous prônons une approche qui prend en compte l'ensemble des données linguistiques disponibles afin de pouvoir repérer les régularités des correspondances graphophonologiques au niveau du système graphique dans sa globalité. Les connaissances sur l'histoire de l'écriture chinoise couplées aux recherches en sciences cognitives, nous permettent de proposer une modélisation de la fiabilité des indices phonologiques que peuvent constituer les composants graphiques au sein de la structure du sinogramme. À court terme, ce modèle sera utilisé pour fournir les outils permettant aux apprenants d'être mieux à même de prédire la prononciation de sinogrammes qu'ils n'ont pas encore vus. L'objectif est de les sensibiliser aux régularités du système graphique au-delà des listes institutionnelles fermées à mémoriser.

Ce modèle est soutenu par une grande quantité de données libres d'origines diverses que nous agrégeons et rediffusons avec une implémentation des calculs proposés. Ainsi, notre contribution ne se limite pas à une liste de composants phonophoriques mais se veut une base théorique et technique pour des outils pédagogiques d'aide à la construction de matériaux ou de suivi des parcours. Le fait de présenter un modèle de calcul plutôt qu'un simple classement, ainsi que la disponibilité des données et du code rendent possibles la prise en compte de la multiplicité des profils, jusqu'au suivi individuel des apprenants.

Dans ce sens, nous travaillons à l'édition d'un volume présentant les phonophores aux apprenants des langues sinogrammiques et pas seulement à ceux qui apprennent le mandarin comme nous nous sommes limités à le faire dans cet article. Cet ouvrage s'accompagnera d'un site Internet et visera à permettre une exploration en autono-

11. On fait ici l'hypothèse que la totalité des composants impliqués dans la liste de sinogrammes est correctement identifiée par l'apprenant.

mie des observations faites par l'intermédiaire de notre modèle. L'objectif est de venir compléter l'usage des méthodes de langues existantes. Par ailleurs, notre contribution, et au-delà notre réseau de connaissances, auraient tout à fait vocation à être intégrés à l'instance qui prend en charge le traitement du lexique au sein d'une plate-forme LMS spécialisée dans l'apprentissage des langues (Mangeot *et al.*, 2016).

Le modèle proposé, fortement motivé par les récents travaux en psycholinguistique, propose une mesure de fiabilité de l'indice phonologique continue plus subtile que celles retenues précédemment par Lee *et al.* (2010) et Hsu *et al.* (2014). Il est aussi plus adapté à un contexte fortement multilingue. Cependant, il reste perfectible. Par exemple, les découpages en trois arguments de l'IDS correspondent assez mal à l'histoire de l'écriture, et certaines interactions entre phonèmes proches d'un point de vue articulatoire ne sont pas capturées de façon pleinement satisfaisantes. Ce sont là deux des points que nous devons approfondir pour améliorer notre modélisation. L'ouverture des données et du code ont aussi pour objectif de faciliter l'amélioration du modèle.

Remerciements

Les auteurs remercient vivement Ilaine Wang et tout particulièrement Guillaume Lechien pour leur aide substantielle lors de la relecture de cet article. Merci également à nos relecteurs pour leur intérêt pour notre travail et leurs remarques constructives. Enfin, merci à Béatrice Pelletier pour la rigueur et la bienveillance de sa relecture formelle.

Une partie du travail présenté dans cet article a été effectuée lors d'un séjour post-doctoral à la l'université nationale *Cheng Kung*, Tainan, financé par la *Taiwan Fellowship* du ministère des Affaires étrangères taïwanais (MOFA).

7. Bibliographie

Allanic B., *La voie des signes : l'apprentissage de la lecture en Chine*, Presses Universitaires de Rennes, Rennes, 2017.

Balibar R., *Le colinguisme*, PUF, impr. 1993, Paris, 1993.

Baxter W. H., Sagart L., *Old Chinese : a new reconstruction*, Oxford University Press, Oxford New York, 2014.

Bellassen J., « La didactique du chinois et la malédiction de Babel. émergence, dynamique et structuration d'une discipline », *Études chinoises,* p. 27-44, 2010.

Boltz W. G., *The origin and early development of the Chinese writing system*, nº 78 in *American oriental series*, American Oriental Soc, New Haven, Conn, 1994.

DeFrancis J., *Visible speech : the diverse oneness of writing systems*, University of Hawaii Press, Honolulu, 1989.

Ding G., Peng D., Taft M., « The Nature of the Mental Representation of Radicals in Chinese : A Priming Study », *Journal of Experimental Psychology : Learning, Memory, and Cognition*, vol. 30, nº 2, p. 530-539, 2004.

Fabre M., « Dyslexie en langue chinoise aujourd'hui. », *Mémoire Master sous la direction de Franck Ramus et Dan XU*, 2009.

Goudin Y., L'intercompréhension entre langues sinogrammiques : représentations, théories, enjeux et applications d'une didactique de la variation, PhD thesis, INALCO, Paris, à paraître.

Goudin Y., Le T. M., « Jouer avec le sacré : les sinogrammes à l'heure des jeux sérieux », *Recherches et applications*, nº 59, p. 145-160, 2016.

Gurevych I., Eckle-Kohler J., Matuschek M., *Linked Lexical Knowledge Bases : Foundations and Applications*, Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers, July, 2016.

Ho C., Chan D. W., Chung K. K., Lee S., Tsang S., « In search of subtypes of Chinese developmental dyslexia », *Journal of Exp. Child Psychology*, nº 97, p. 61-83, 2007.

Ho C., Chan D. W., Lee S., Tsang S., Luan V., « Cognitive profiling and preliminary subtyping in Chinese developmental dyslexia », *Cognition*, nº 91, p. 43-75, 2004.

Ho C., Law T. P., Ng P., « The phonological deficit hypothesis in Chinese developmental dyslexia », *Reading and Writing*, nº 13, p. 57-79, 2000.

Ho C. S., Bryant P., « Learning to read Chinese beyond the logographic phase », *Reading Research Quarterly*, nº 32, p. 276-289, 1997a.

Ho C. S., Bryant P., « Phonological skills are important in learning to read Chinese », *Developmental Psychology*, nº 33, p. 946-951, 1997b.

Hsu C., Lee C., Tzeng O., « Early MEG markers for reading Chinese phonograms : Evidence from radical combinability and consistency effects », *Brain and Language*, vol. 139, p. 1-9, December, 2014.

Lee C., Huang H., Kuo W., Tsai J., Tzeng O., « Cognitive and neural basis of the consistency and lexicality effects in reading Chinese », *Journal of Neurolinguistics*, nº 23, p. 10-27, 2010.

Lyssenko N., Anastasia, *Méthode programmée du chinois moderne*, Paris, 1986.

Magistry P., Goudin Y., Wang I., Lechien G., « Building a network of knowledge on sinograms », *28e Journées de Linguistique d'Asie Orientale (JLAO)*, 2015.

Mangeot M., Bellynck V., Eggers E., Loiseau M., Goudin Y., « Exploitation d'une base lexicale dans le cadre de la conception de l'ENPA Innovalangues », *in* I. S. et Jovan Kostov (ed.), *Enseignement des Langues et TAL*, vol. 09 of *ELTAL*, ATALA et AFCP, Paris, France, p. 48-64, 2016.

McBride-Chang C., Ho C. S., « Developmental issues in Chinese children's character acquisition », *Journal of Educational Psychology*, vol. 92, nº 1, p. 50-55, 2000.

Prevot L., Magistry P., Huang C.-R., « Un état des lieux du traitement automatique du Chinois », *Faits de langues*, vol. 46, nº 2, p. 61-70, 2015.

Shannon C. E., « A mathematical theory of communication », *The Bell System Technical Journal*, 1948.

Shu H., Anderson R., « Learning to read Chinese : The development of metalinguistic awareness », *J. Wang. AW. Inhoff et HC, Chen (Eds.), Reading Chinese script- A cognitive analysis. Mahwah. NJ : Lawrence Erlbaum Ass. Inc.*p. 1-18, 1999.

Shu H., Chen X., Anderson R., Wu N., Xuan Y., « Properties of school Chinese : Implications for learning to read », *Child Development*, nº 74, p. 27-47, 2003.

Siok W., Fletcher P., « The role of phonological awareness and visual-orthographic skills in Chinese reading acquisition », *Developmental Psychology*, vol. 37, nº 6, p. 886-899, 2001.

Wang C., Peng D., « Basic processing unit of Chinese character recognition : Evidence from stroke number effect and radical number effect. », *Acta Psychologica Sinica*, 1997.

Windfuhr K., Snowling M., « The relationship between paired associate learning and phonological skills in normally developing readers », *Journal of Experimental Child Psychology*, vol. 80, nº 2, p. 160-173, 2001.

Wydell T., Butterworth B., « A case study of an English-Japanese bilingual with monolingual dyslexia », *Cognition*, vol. 70, nº 3, p. 273-305, April, 1999.

Ziegler J., Goswami U., « Reading acquisition, developmental dyslexia, and skilled reading across languages : A psycholinguistic grain size theory », *Psychological Bulletin*, nº 131, p. 3-29, 2005.