Documentation et bibliothèques

DOCUMENTATION BIBLIOTHÈQUES

Problèmes de repérage des ressources bibliographiques en langue chinoise : une perspective occidentale
On the Retrieval of Bibliographic Resources in Chinese: A Western Perspective
Problemas de localización de fuentes bibliográficas en chino

Clément Arsenault

Volume 51, Number 3, July–September 2005

URI: https://id.erudit.org/iderudit/1029496ar
DOI: https://doi.org/10.7202/1029496ar

See table of contents

Publisher(s)

Association pour l'avancement des sciences et des techniques de la documentation (ASTED)

ISSN

0315-2340 (print)
2291-8949 (digital)

Explore this journal

Cite this article

Arsenault, C. (2005). Problèmes de repérage des ressources bibliographiques en langue chinoise : une perspective occidentale. *Documentation et bibliothèques*, *51*(3), 175–184. https://doi.org/10.7202/1029496ar

Article abstract

This article outlines several research possibilities regarding the development of retrieval methods for bibliographic records in Chinese. Past research has underscored the successes and limits of retrieval based on Romanized Chinese data (Pinyin). It would appear that a significant proportion of users do not obtain satisfactory results when searching in Pinyin. In order to provide these users with better retrieval methods, it is essential to explore other options that can be integrated to the bibliographic data bases in a North American context, where the documents in Chinese usually make up a small portion of the collections.

DOCUMENTATION BIBLIOTHÈQUES

Problèmes de repérage des ressources bibliographiques en langue chinoise : une perspective occidentale*

CLÉMENT ARSENAULT
Clement.arsenault@umontreal.ca

RÉSUMÉ | ABSTRACTS | RESUMEN

Ce travail a pour objet de présenter plusieurs avenues de recherche pour le développement de modules de repérage de notices bibliographiques en langue chinoise. Des études antérieures ont montré les succès et les limites du repérage se fondant sur des données chinoises romanisées (pinyin). Il semble qu'une proportion non négligeable des utilisateurs n'obtient pas des résultats très satisfaisants lors du repérage en pinyin. Pour fournir à ces utilisateurs des moyens de repérage mieux adaptés, il est essentiel d'explorer d'autres avenues méthodologiques susceptibles d'être intégrées aux bases bibliographiques dans le contexte nord-américain, où les ressources en langue chinoise représentent habituellement seulement une proportion minime des collections..

On the Retrieval of Bibliographic Resources in Chinese: A Western Perspective*

This article outlines several research possibilities regarding the development of retrieval methods for bibliographic records in Chinese. Past research has underscored the successes and limits of retrieval based on Romanized Chinese data (Pinyin). It would appear that a significant proportion of users do not obtain satisfactory results when searching in Pinyin. In order to provide these users with better retrieval methods, it is essential to explore other options that can be integrated to the bibliographic data bases in a North American context, where the documents in Chinese usually make up a small portion of the collections.

Problemas de localización de fuentes bibliográficas en chino*

El objetivo de este trabajo es presentar varias alternativas de investigación para el desarrollo de módulos de localización de reseñas bibliográficas en chino. Estudios anteriores mostraron los aciertos y desaciertos de la localización basados en datos chinos transliterados en alfabeto latino (sistema pinyin). Al parecer un número considerable de usuarios no obtiene resultados muy satisfactorios cuando busca información bibliográfica con este sistema. Por este motivo y con el fin de facilitarles medios de localización mejor adaptados, es indispensable explorar otras posibilidades metodológicas capaces de integrarse a las bases bibliográficas del contexto norteamericano en las que las fuentes en chino representan normalmente una proporción de las colecciones.

* Cette recherche a été rendue possible grâce à une subvention du Conseil de recherches en sciences humaines du Canada. Ce texte a été publié originalement en anglais dans *International Information and Library Review*.

* *This research was made possible with a grant from the Social Sciences and Humanities Research Council of Canada. This article was originally published in English in the* International Information and Library Review.

* Esta investigación se realizó gracias a la subvención del Consejo de Investigación en Ciencias Humanas de Canadá. La publicación original del texto fue en inglés, en la revista *"Intenational Information and Library Review"*.

CONTEXTE DE LA RECHERCHE

Dans un environnement où l'information est enregistrée principalement en caractères romains, le repérage d'information textuelle en langue chinoise présente des défis particuliers et spécifiques que les systèmes conventionnels ne relèvent pas avec une efficacité souhaitable. Dans le monde occidental, les systèmes de repérage informatisés sont habituellement conçus en fonction des besoins de repérage de ressources enregistrées dans les langues occidentales et sont, par conséquent, mal adaptés au repérage de ressources en langue chinoise.

Dans les systèmes en ligne, le repérage de données en langue chinoise se fait ordinairement soit en lançant des requêtes écrites en vernaculaire à la recherche de correspondances en vernaculaire parmi les entrées indexées, soit en lançant des requêtes romanisées à la recherche de correspondances parmi les entrées romanisées des index.

Du point de vue de l'utilisateur, les deux méthodes soulèvent des problèmes et des obstacles spécifiques. Dans le premier cas, générer des caractères chinois pour la formulation de sa requête peut présenter un défi à l'utilisateur. L'utilisation des systèmes d'entrée de données conventionnels tels que le clavier d'ordinateur est loin d'être une façon idéale de générer des caractères chinois même si, dans les années récentes, des « éditeurs de méthodes d'entrée » (*input method editors* ou IME) ont été développés et intégrés aux systèmes d'exploitation tels que Microsoft Windows et ont facilité cette tâche. Dans le second cas, lorsque le repérage se fait sur des entrées romanisées, le problème rencontré est celui d'un niveau élevé d'homonymie, qui dilue la précision des résultats du repérage. Cet obstacle est dû au fait que les marqueurs de ton ne sont généralement pas enregistrés ou pris en compte dans le processus de l'indexation ; il est dû aussi au fait que le texte est converti en unités lexicales monosyllabiques au lieu de polysyllabiques, comme c'est le cas, par exemple, des champs romanisés des notices bibliographiques MARC. À l'heure actuelle, le système de romanisation le plus utilisé pour la transcription des caractères chinois dans les notices bibliographiques est le système pinyin, qui a été développé

en Chine au milieu des années 1950. Les grandes bases de données bibliographiques, telles que l'OCLC (*Online Computer Library Center*) et le RLG (*Research Libraries Group*), contiennent des notices en caractères pinyin. En avril 2005, on comptait plus de 1,32 million de notices en langue chinoise dans la base WorldCat de l'OCLC (OCLC, 2005). Il est cependant important de noter que l'information tonale n'est pas enregistrée. Il y a quatre tons distincts dans le chinois standard moderne, ainsi qu'un ton neutre.

Une récente recherche effectuée par Tull (2002) a montré qu'un bon nombre de systèmes automatisés de bibliothèque ont déjà implanté la norme Unicode à des degrés divers, mais que, pour diverses raisons, peu de bibliothèques nord-américaines offrent pour l'instant un système de repérage où il soit possible de faire une recherche sur les données vernaculaires. Traditionnellement, les bibliothèques s'appuient sur des textes romanisés pour la recherche, le tri et l'affichage des notices bibliographiques de matériaux en langue chinoise. Il est donc encourageant de constater l'émergence récente d'un nombre grandissant de catalogues à accès public en ligne (*online public access catalogs* ou OPACs) munis de fonctionnalités multiscript. Certains systèmes offrent la possibilité d'afficher des caractères vernaculaires chinois et de soumettre des requêtes en chinois vernaculaire[1], tandis que d'autres permettent l'affichage de caractères non romains, mais n'ont pas encore la capacité de traiter des requêtes formulées en caractères non romains[2].

Des études récentes ont montré qu'en ce qui concerne le chinois, la romanisation permet des recherches assez efficaces dans les titres de monographies (Arsenault, 2000; Mair, 2001). L'étude effectuée par Huang sur le catalogue de l'Université de Pékin révèle que le repérage en pinyin peut être amélioré et facilité si le système impose les conditions de séquence et de contiguïté tout en ignorant automatiquement les espaces entre les termes (Huang, 2004). Mais il est fort improbable que ces choix puissent être programmés comme paramètres par défaut dans des systèmes de repérage où les notices en pinyin sont minoritaires. Une étude récente a montré que l'utilisation de la romanisation (pinyin dans ce cas) pour le repérage donne de bons résultats, de l'avis d'une large portion des utilisateurs. Mais il est intéressant d'observer qu'une proportion non négligeable des participants, qui tous avaient affirmé leur familiarité avec le pinyin, ont eu des difficultés à compléter une tâche de repérage simple en pinyin. En réalité, plusieurs facteurs, dont le niveau d'éducation et l'interférence dialectale, affectent profondément le niveau de familiarité des utilisateurs avec le pinyin et leur évaluation du pinyin dans la recherche d'éléments en langue chinoise dans un catalogue public en ligne. Pour fournir à cette clientèle des services de repérage plus adéquats et plus efficaces, il est donc essentiel d'examiner les options méthodologiques en tenant compte des disparités dans la familiarité avec le pinyin au sein des sous-groupes d'utilisateurs. Il apparaît essentiel d'adapter un certain nombre de techniques de repérage selon les besoins spécifiques de chaque requête et de chaque utilisateur. Il est également nécessaire de s'interroger sur l'applicabilité et l'adaptabilité de ces techniques de repérage dans un environnement nord-américain.

La présente étude a pour but d'examiner les façons possibles d'intégrer une variété de modules de repérage dans les grands OPACs multilingues accessibles via Internet, afin de repérer les objets en langue chinoise qui y sont catalogués. Les problèmes reliés au repérage des notices bibliographiques en langue chinoise en contexte nord-américain sont abordés. Dans le but de faciliter le repérage de documents en langue chinoise dans les catalogues à accès public en ligne, notre étude suggère et présente plusieurs avenues de recherche sur ce thème.

Recherches antérieures

Dans une recherche antérieure, nous avons rassemblé des données en vue de mesurer l'efficacité et le rendement du pinyin dans le repérage des titres chinois dans les OPACs. L'analyse des données a montré que le taux de succès est assez élevé dans le cas des recherches sur des objets spécifiques. Vingt-quatre participants ayant tous le chinois pour langue maternelle ont eu à réaliser des recherches sur des objets spécifiques (c'est-à-dire sur des titres spécifiques plutôt que sur des thèmes donnés), sur 40 titres chinois, en utilisant le pinyin dans un gros catalogue public en ligne. Le taux de succès constaté au cours de cette expérience se situe entre 80 % et 90 %, selon le modèle d'agrégation et le mode de recherche utilisés par les participants (Arsenault, 2000 : 154). Les entrées en pinyin pouvaient être enregistrées selon un modèle monosyllabique (non agrégé), ou selon un modèle polysyllabique (agrégé), suivant les politiques de développement locales. En raison de la petite taille de l'échantillonnage, les variations entre les groupes n'ont pas été jugées significatives du point de vue statistique. Notons en passant que le succès dans les recherches sur des objets spécifiques se définit comme le fait de trouver, c'est-à-dire d'afficher la notice bibliographique de l'objet recherché.

En dépit de ce taux de succès plutôt élevé, il est intéressant de noter qu'approximativement la moitié des requêtes formulées par les participants ont échoué,

1. Parmi ceux qui valent la peine d'être mentionnés, citons l'OPAC de la University of California (MELVYL) <*http://melvyl.cdlib.org*>, celui de Harvard University (Hollis) <*http://holliscatalog.harvard.edu*>, ainsi que celui de la University of Massachusetts <*http://fclibr.library.umass.edu*>.
2. C'est le cas du catalogue WorldCat de l'OCLC sur l'interface FirstSearch <*http://firstsearch.oclc.org*>, du catalogue de la University of British Columbia <*http://webcat.library.ubc.ca*>, et de celui de Yale University (Orbis) <*http://orbis.library.yale.edu*>, pour ne nommer que ceux-là.

TABLEAU 1 :
Taux moyen de succès de la première requête dans les recherches en pinyin sur des objets spécifiques dans un OPAC.

	PINYIN (MONOSYLLABES)	PINYIN (POLYSYLLABES)
Titre exact	59% (*n*=12)	57% (*n*=12)
Mots-clés dans le titre (*Keywords-in-title*)	47% (*n*=12)	48% (*n*=11)

ce qui signifie que pour chaque objet recherché, approximativement deux requêtes ont été nécessaires. Le taux de succès tombe radicalement si l'on ne considère que la première requête (voir le tableau 1).

Les données ayant été rassemblées dans un contexte expérimental, on peut supposer que les participants, se sachant observés, avaient à cœur de produire des résultats et n'hésitaient pas à reformuler leur requête afin de repérer la notice, même s'il n'y avait aucune garantie que la notice se trouve effectivement dans la base de données. Dans une situation de la vie courante, il se pourrait que les utilisateurs soient moins motivés à répéter leur recherche plusieurs fois. Par conséquent, les taux de succès pourraient être moins élevés que ceux mesurés dans le contexte expérimental. Mais on pourrait aussi trouver des arguments en faveur de la position opposée : les participants sont plus motivés dans une situation de la vie courante, parce qu'ils ont un besoin réel de l'information qu'ils recherchent. Les recherches sur des titres spécifiques sont les opérations de repérage les plus faciles qu'on puisse faire sur un OPAC, plus faciles, certainement, que les recherches par sujet. Pourtant, il est intéressant de constater que les participants ont eu de la difficulté à localiser les notices. Plusieurs sources d'erreur ont été identifiées qui expliquent les échecs. La majorité peut être attribuée à des erreurs d'agrégation ou à des erreurs de romanisation. Le tableau 2 donne le détail de ces erreurs dans trois catégories : (1) la proportion des erreurs que l'on peut attribuer à une non correspondance entre le modèle d'agrégation de la requête et le modèle d'agrégation des entrées d'indexation ; (2) la proportion qui peut être attribuée aux erreurs de romanisation, et (3) les autres erreurs.

En moyenne, chacun des 24 participants a fait 65 erreurs sur les 40 titres qui faisaient l'objet de recherches. Il faut néanmoins noter que la distribution des erreurs de romanisation parmi les participants présente une structure bimodale très marquée : une large proportion des participants fait peu d'erreurs, tandis qu'une proportion plus faible, mais tout de même importante, fait un grand nombre d'erreurs (Arsenault, 2002 : 49), alors que les erreurs d'agrégation sont distribuées de manière plus uniforme. Il

TABLEAU 2 :
Nombre d'erreurs observées dans les requêtes ayant échoué.

	PINYIN (MONOSYLLABES)	PINYIN (POLYSYLLABES)
Erreurs d'agrégation	308 (43,7%)	494 (56,9%)
Erreurs de romanisation	348 (49,4%)	319 (36,8%)
Autres erreurs	49 (6,9%)	55 (6,3%)
Total	705 (100%)	868 (100%)

apparaît donc qu'une proportion non négligeable des utilisateurs n'est pas parfaitement à l'aise avec le repérage selon la romanisation, qui est fréquemment la seule méthode offerte par les OPACs et certains autres systèmes de repérage en ligne.

Cette analyse nous amène à conclure que l'utilisation de la romanisation comme seul et unique moyen de repérer des titres spécifiques ne produit pas des résultats satisfaisants. D'autres options méthodologiques sont nécessaires si l'on désire améliorer le repérage et la performance des utilisateurs. De plus, il serait souhaitable d'incorporer des techniques de repérage qui prennent en compte le fait que l'agrégation des syllabes chinoises en unité lexicales est fréquemment ambiguë, du fait qu'il n'existe pas de standard orthographique largement accepté et bien promu dans ce domaine (Yin et Felley, 1990). Il n'y a aucune garantie que l'utilisateur sera en mesure de deviner comment une chaîne de caractères chinois a été agrégée dans sa forme romanisée, comme le montre clairement le fait que les erreurs d'agrégation comptent pour la moitié de toutes les erreurs (tableau 2).

Ces observations nous mènent à proposer les aires de recherche suivantes, dans le but d'améliorer la qualité du repérage d'objets en langue chinoise :

- ◊ L'identification des requêtes par traitement linguistique informatisé ;
- ◊ Le repérage fondé sur la pertinence dans les recherches sur des titres spécifiques ;
- ◊ Le repérage trans-scripts.

LE TRAITEMENT LINGUISTIQUE INFORMATISÉ

On peut imaginer que le catalogue public en ligne d'une collection contenant une majorité d'ouvrages en langue chinoise (par exemple, le catalogue d'un département d'études de l'Asie de l'Est dans une grande institution universitaire) pourrait offrir une interface utilisateurs permettant de sélectionner le mode de repérage : l'utilisateur pourrait choisir entre le repérage selon la romanisation, le repérage selon des requêtes

TABLEAU 3 :
Résultats d'une expérience d'identification de la langue en utilisant un court extrait de texte chinois.

DÉVELOPPEUR	PRODUIT	URL	VERNACULAIRE[a]	PINYIN NON AGRÉGÉ[b]	PINYIN AGRÉGÉ[c]
Xerox	CA	<www.xrce.xerox.com/competencies/content-analysis/tools/guesser.en.html>	Chinois	Catalan	Catalan
Lextex	Lextex Intl.	<www.lextek.com>	Abkhaz	Javanais	Javanais
Alfa-informatica, U. de Groningen	Textcat	<odur.let.rug.nl/~vannoord/TextCat/Demo/textcat.html>	Chinois	Inconnu	Inconnu
RALI, U. de Montréal	SILC	<www-rali.iro.umontreal.ca/SILC/SILC.fr.cgi>	Chinois	Allemand	Italien
Alis Technologies	¿Qué?[4]	<quebec.alis.com/castil/essai_silc.cgi>	Chinois	Allemand	Italien
PentaMem Technology	PentaMem	<nlp.petamem.com/langrec.cgi>	Albanais	Turc	Turc
MorphoLogic	LangWitch	<www.morphologic.hu/order/langwitch.asp>	Pas disp.	Espagnol	Espagnol

a) 在过去35年里，日本皇室的9名新生儿全是女的，严重缺乏男性继位人选。日本国会被迫考虑修改宪法，以接纳一位女天皇
b) Zai guo qu 35 nian li, Ri ben huang shi de 9 ming xin sheng er quan shi nü de, yan zhong que fa nan xing ji wei ren xuan. Ri ben guo hui bei po kao lü xiu gai xian fa, yi jie na yi wei nü tian huang.
c) Zai guoqu 35 nian li, Riben huangshi de 9 ming xinsheng'er quan shi nü de, yanzhong quefa nanxing jiwei renxuan. Riben guohui beipo kaolü xiugai xianfa, yi jiena yi wei nü tianhuang.

formulées en caractères chinois, ou une combinaison des deux. Toutefois cette approche pourrait s'avérer peu pratique si la collection chinoise est intégrée à un ensemble multilingue plus grand, destiné à une clientèle diverse, et comportant un vaste assortiment de collections.

Une solution possible consisterait à développer des agents intelligents et à les intégrer à l'interface du système de repérage. Ces agents se définissent comme des modules informatiques semi-autonomes capables d'identifier des modèles répétitifs de comportements, des similitudes entre des événements et des objets, et des changements de modèles dans le temps (Feldman et Yu[3]). Ces agents pourraient agir sur plusieurs fronts pour faciliter le repérage d'objets chinois. À l'aide de techniques d'identification de la langue lors de l'analyse des termes de la requête, un agent pourrait détecter que l'utilisateur est à la recherche d'un objet en chinois. Cette détection achevée, l'agent intelligent pourrait afficher une interface offrant un assortiment de techniques de repérage adaptées au repérage de notices d'ouvrages en langue chinoise, puis cibler automatiquement le sous-ensemble des notices bibliographiques en langue chinoise contenues dans le catalogue. En cas d'échec de la requête ou d'un manque de précision, l'agent pourrait activer et adapter des algorithmes de pertinence qui pourraient classer par ordre de pertinence de grands ensembles de notices ou appeler des techniques d'expansion de requêtes. Enfin, la détection de la langue par l'agent pourrait aussi déclencher l'activation de modules de repérage trans-scripts qui pourraient comparer les termes des requêtes formulées en vernaculaire aux termes d'indexation romanisés et l'inverse.

Dans ce cadre de travail, il est évidemment essentiel que la langue de la requête soit identifiée avant l'activation de toute autre technique de repérage. Idéalement cela devrait être fait en demandant simplement à l'utilisateur d'identifier manuellement la langue de sa recherche. En pratique, cependant, il y a peu de chances d'obtenir une telle information des utilisateurs, qui n'exploitent que rarement les options avancées de repérage (Xie et O'Hallaron, 2002 : 2). La plupart des études sur l'identification informatique de la langue se concentrent sur l'analyse statistique selon les caractéristiques textuelles. Dans le cas des textes vernaculaires chinois, l'opération est assez simple, mais l'identification informatique de la langue d'un texte chinois romanisé pourrait être plus difficile, surtout quand l'analyse opère sur les quelques termes d'une requête. Il existe un certain nombre de produits informatiques commerciaux qui s'attaquent au défi d'identifier la langue d'un texte écrit. Le tableau 3 ci-dessous montre les résultats obtenus lors d'une expérience lancée à partir d'un petit texte provenant du site Internet de Yahoo! News en chinois (http://cn.news.yahoo.com). Quatre des sept produits testés ont été capables d'identifier correctement la langue d'un texte vernaculaire, mais aucun n'a pu l'identifier à partir des deux formes pinyin. Il est donc clair que de nouveaux développements sont requis dans ces produits avant de pouvoir détecter un texte chinois en forme romanisée.

La question de savoir si cette identification linguistique est possible dans le cadre de requêtes dans un catalogue public en ligne (OPAC), et avec quel niveau de succès, est évidemment essentielle. Pour tenter de répondre à cette question, nous avons procédé à une autre brève expérience, dont les résultats sont encourageants. Nous avons utilisé des syllabes pinyin indivi-

3. Notre traduction.

duelles pour formuler des requêtes dans l'OPAC de la Library of Congress <http://catalog.log.gov>, un très vaste catalogue multilingue contenant plus de 12 millions d'objets. Chacune des 407 syllabes[4] a été utilisée individuellement pour lancer des recherches sur des mots-clés dans le titre, une première fois sans limite de langue, une seconde fois avec la langue limitée aux notices en langue chinoise uniquement. On trouvera les données complètes en annexe. Le ratio des deux recherches est ainsi une indication de la probabilité que chaque syllabe pinyin correspond bien à du texte chinois. Par exemple, il est hautement probable que si la syllabe « xiang » est utilisée dans une requête, sa présence indique l'intention de repérer un titre chinois, puisque 98 % de toutes les occurrences de « xiang » dans l'index viennent de notices en langue chinoise. D'autre part, du fait que 1 % seulement des occurrences de la syllabe « run » correspondent à des données chinoises, il est hautement improbable qu'une requête contenant ce terme indique l'intention de repérer un titre en langue chinoise. Ce qu'il y a d'encourageant et de prometteur dans cette brève analyse est que 25 % environ de toutes les syllabes pinyin ont une probabilité de plus de 90 % de correspondre à des notices en langue chinoise. De plus, la probabilité demeure assez élevée (plus de 69 %) si on ne considère que la moitié des syllabes. On peut dès lors imaginer que si une requête contient simplement deux ou trois syllabes, il est assez facile de détecter, dans la majorité des cas et avec un bon niveau de fiabilité, qu'une requête est formulée en pinyin et que, par conséquent, elle indique l'intention de repérer des ressources en langue chinoise.

Le repérage fondé sur la pertinence dans les recherches sur des titres spécifiques

Traditionnellement, c'est aux recherches par sujet que s'appliquent les techniques de repérage fondées sur la pertinence. Dans les recherches sur des titres spécifiques, le taux de succès et le ratio de précision sont habituellement suffisamment élevés pour ne pas opposer d'obstacles à un repérage efficace. Mais comme nous l'avons vu plus haut, le taux de succès d'une première requête sur des titres spécifiques chinois peut ne pas être satisfaisant en raison de l'ambiguïté concernant l'agrégation et de la confusion dans la romanisation, qui sont dues au fait que la langue chinoise contient un volume d'information phonologique très élevé, si l'on prend en compte toutes les variables en jeu. Par conséquent, la moindre nuance est susceptible de faire une grande différence au repérage et la mauvaise prononciation d'un mot va aboutir, selon toute probabilité, à la prononciation d'un autre mot (Chao, 1968 : 23[5]). Une analyse plus poussée des requêtes a permis de regrouper les erreurs d'agrégation et les erreurs de romanisation en trois ou quatre sous-groupes. En ce qui concerne les erreurs de romanisation, nous avons observé que la majorité des requêtes étaient en fait presque des correspondances, la différence étant due à une légère confusion de prononciation chez l'utilisateur (Arsenault, 2000 : 183). Par exemple, le terme « lin », qui signifie « forêt » a parfois été introduit sous la graphie « ling ». La confusion entre les prononciations nasale avant (consonne alvéolaire) et nasale arrière (consonne vélaire) est assez fréquente chez les locuteurs de langue maternelle chinoise (Chao, 1961 : 7 ; King, 1983 : 98-99 ; Yin et Felley, 1990 : 27-28). Parmi les autres erreurs de prononciation observées fréquemment, il faut compter la confusion entre les paires fricatives (s/sh, ch/zh...) et les autres paires de consonnes (l/n, h/f, par exemple).

Ces observations nous amènent à souhaiter que des données plus abondantes soient rassemblées concernant ces phénomènes, et que des algorithmes de repérage prenant en compte les correspondances approximatives soient développés. L'application de techniques de recherches floues pourrait être utile ici, à cause de leur tolérance des erreurs commises tant à l'entrée des données qu'à l'entrée des requêtes (Chu, 2003 : 65). La même chose vaut pour les erreurs d'agrégation, pour lesquelles des algorithmes de classement par proximité de termes pourraient améliorer le repérage même dans le cas de recherches de titres spécifiques.

Le repérage trans-scripts

Le repérage trans-scripts est une autre aire de recherche qui mérite d'être approfondie. Il est très étroitement lié au repérage d'information multilingue (RIML), qui se définit comme la découverte de documents dans une langue répondant à des requêtes formulées dans une autre langue (Xu, Weischedel et Nguyen, 2001 : 105[6]). Comme nous l'avons mentionné au début de cette étude, le repérage des données en langue chinoise se fait traditionnellement en cherchant une ou des correspondances entre des termes de requête romanisés et des données romanisées, ou bien entre des requêtes en vernaculaire et des données vernaculaires (voir la figure 1). Notons à nouveau que la plupart des OPACs du monde occidental se limitent au repérage selon la romanisation.

Pour le repérage de notices contenant à la fois des entrées vernaculaires et romanisées, il devrait être possible d'appliquer des techniques RIML. Il

4. Il y a 409 syllabes en tout, mais comme le système ignore les signes diacritiques dans les termes de requête, les paires syllabiques lu/lü et nu/nü ont dû être confondues.

5. Notre traduction.

6. Notre traduction.

FIGURE 1 :
Modèle de repérage traditionnel.

ENTRÉE	TERMES D'INTERROGATION	TERMES D'INDEXATION
Romanisation →	Romanisation →	Romanisation
Autres méthodes d'entrée →	Vernaculaire →	Vernaculaire

FIGURE 2 :
Modèle de repérage trans-scripts (cross-script retrieval).

ENTRÉE	TERMES D'INTERROGATION	TERMES D'INDEXATION
Romanisation →	Romanisation →	Romanisation
Autres méthodes d'entrée →	Vernaculaire →	Vernaculaire

faudrait seulement dans ce cas appliquer le repérage entre scripts plutôt qu'entre langues. Ainsi le repérage ne serait plus limité aux seules recherches de correspondance entre requêtes romanisées et données romanisées, il pourrait se faire selon plusieurs avenues, comme le montre la figure 2 ci-dessous.

Dans ce modèle, l'utilisateur serait libre de formuler sa requête sous n'importe quelle forme, et il serait possible d'y incorporer des techniques de repérage et d'indexation développées spécifiquement pour du texte vernaculaire chinois, basées sur les caractères ou sur les mots, ou encore des méthodes hybrides telles que les méthodes *n*-grams (Nie et Ren, 1999 ; Dai, Khoo et Loh, 1999).

IMPLICATIONS DE LA RECHERCHE

Nous avons présenté, dans cette étude, la problématique générale du repérage de titres spécifiques en langue chinoise dans les catalogues publics en ligne nord-américains (OPACs). La recherche de titres spécifiques dans un OPAC est une opération ordinairement facile, mais elle devient fréquemment complexe quand il s'agit d'une recherche d'objets en langue chinoise, surtout quand les modules d'interrogation et de repérage ne sont pas adaptés aux particularités de cette langue. Notre analyse de quelques-uns des problèmes que pose cette situation montre qu'il existe un besoin urgent de recherches dans ce domaine, si l'on veut améliorer et faciliter le repérage d'objets en langue chinoise. Le but de la présente étude a été d'identifier des avenues de recherche possibles et de mettre de l'avant quelques idées sur la manière de les mener à bien. ◉

SOURCES CONSULTÉES

Arsenault, Clément. 2002. Pinyin Romanization for OPAC retrieval: is everyone being served? *Information Technology and Libraries*, 21(2): 45-50.

_____. 2000. *Word division in the transcription of Chinese script in the title fields of bibliographic records.* Thèse de doctorat non publiée. Toronto, Université de Toronto. Disponible chez UMI, n° AAT NQ53736.

Chao, Yuen Ren. 1961. *Mandarin primer: an intensive course in spoken Chinese.* Cambridge, Harvard University Press.

_____. 1968. *A grammar of spoken Chinese.* Berkeley, University of California Press.

Chu, Heting. 2003. Information representation and retrieval in the digital age. Medford, N.J., *Information Today.*

Dai, Yubin, Christopher S.G. Khoo et Teck Ee Loh. 1999. A new statistical formula for Chinese text segmentation incorporating contextual information. *SIGIR '99:* 82-89.

Feldman, Susan et Edmund Yu. 1999. Intelligent agents: a primer. *Searcher*, 7 (9).

Huang, Jie. 2004. Retrieval of Chinese language in pinyin: a comparative study. *Information Technology and Libraries*, 23 (3): 95-100.

King, Paul L. 1983. *Contextual factors in Chinese pinyin writing.* Thèse de doctorat non publiée. Ithaca, Cornell University.

Mair, Victor H. 2001. Pinyin orthographical rules for libraries: a follow-up. *Chinese Librarianship. An International Electronic Journal*, 7.

Nie, Jian-Yun et Fuji Ren. 1999. Chinese information retrieval: Using characters or words? *Information Processing & Management*, 35 (4): 443-462.

OCLC. 2005. WorldCat facts and statistics. <http://www.oclc.org/worldcat/statistics/> (page consultée le 2 mai 2005).

Tull, Laura. 2002. Library systems and Unicode: a review of the current state of development. *Information Technology and Libraries*, 21 (4): 181-185.

Xie, Yinglian et David O'Hallaron. 2002. Locality in search engine queries and its implications for caching. *Proceedings of IEEE INFOCOM — The Conference on Computer Communications.* <http://www-2.cs.cmu.edu/~droh/papers/queryinfocom.pdf>.

Xu, Jinxi, Ralph Weischedel et Chanh Nguyen. 2001. Evaluating a probabilistic model for cross-lingual information retrieval. *SIGIR '01:* 105-110.

Yin, Binyong et Mary Felley. 1990. *Chinese Romanization: pronunciation and orthography.* Beijing, Sinologua.

ANNEXE

Données assemblées à partir du catalogue en ligne de la Library of Congress entre le 8 novembre 2002 et le 30 janvier 2003.

Syllabe	KW ds Titre	KW ds Titre & la=Chinois	Ratio	Rang	Rang cum.
qiong	86	86	100,00 %	1	0,25 %
zhei	1	1	100,00 %	2	0,49 %
zhui	378	377	99,74 %	3	0,74 %
ceng	576	573	99,48 %	4	0,98 %
yu	*37601*[5]	*37323*	99,26 %	5	1,23 %
xian	16822	16646	98,95 %	6	1,47 %
jiao	12398	12263	98,91 %	7	1,72 %
jia	10745	10619	98,83 %	8	1,97 %
bian	10575	10446	98,78 %	9	2,21 %
zuo	9398	9283	98,78 %	10	2,46 %
xiu	2285	2257	98,77 %	11	2,70 %
qiao	949	937	98,74 %	12	2,95 %
xue	36015	35505	98,58 %	13	3,19 %
xiong	656	646	98,48 %	14	3,44 %
xin	14323	14097	98,42 %	15	3,69 %
dian	11685	11499	98,41 %	16	3,93 %
jian	15370	15114	98,33 %	17	4,18 %
nian	11729	11526	98,27 %	18	4,42 %
xing	10067	9891	98,25 %	19	4,67 %
xiao	10638	10452	98,25 %	20	4,91 %
jiu	17008	16699	98,18 %	21	5,16 %
zhong	15016	14733	98,12 %	22	5,41 %
zhun	530	520	98,11 %	23	5,65 %
qi	13818	13551	98,07 %	24	5,90 %
zong	4012	3934	98,06 %	25	6,14 %
xiang	10150	9949	98,02 %	26	6,39 %
qian	4053	3970	97,95 %	27	6,63 %
zheng	12095	11847	97,95 %	28	6,88 %
jue	1597	1564	97,93 %	29	7,13 %
zhuang	1402	1373	97,93 %	30	7,37 %
guan	11272	11038	97,92 %	31	7,62 %
xia	2597	2543	97,92 %	32	7,86 %
guo	15220	14897	97,88 %	33	8,11 %
jie	10923	10690	97,87 %	34	8,35 %
zhai	1140	1115	97,81 %	35	8,60 %
qiang	960	938	97,71 %	36	8,85 %
jiang	4974	4857	97,65 %	37	9,09 %
zhuan	9561	9336	97,65 %	38	9,34 %
zhu	13807	13474	97,59 %	39	9,58 %
qing	9002	8782	97,56 %	40	9,83 %
ren	14828	14464	97,55 %	41	10,07 %
guang	2075	2021	97,40 %	42	10,32 %
cang	1689	1645	97,39 %	43	10,57 %
zhan	8556	8328	97,34 %	44	10,81 %
diao	1834	1785	97,33 %	45	11,06 %
zhuo	182	177	97,25 %	46	11,30 %
zhen	4227	4110	97,23 %	47	11,55 %
qu	7327	7119	97,16 %	48	11,79 %
xun	3019	2933	97,15 %	49	12,04 %
lian	2625	2550	97,14 %	50	12,29 %
zha	420	408	97,14 %	51	12,53 %
shuai	388	376	96,91 %	52	12,78 %
qiu	2148	2078	96,74 %	53	13,02 %
zhua	30	29	96,67 %	54	13,27 %
cun	2302	2225	96,66 %	55	13,51 %
xuan	12568	12145	96,63 %	56	13,76 %
gou	1691	1634	96,63 %	57	14,00 %
yue	3738	3611	96,60 %	58	14,25 %
gong	15218	14696	96,57 %	59	14,50 %
qun	611	590	96,56 %	60	14,74 %
zi	15689	15143	96,52 %	61	14,99 %
cuo	226	218	96,46 %	62	15,23 %
biao	2271	2190	96,43 %	63	15,48 %
zhang	3293	3171	96,30 %	64	15,72 %
zhao	1480	1424	96,22 %	65	15,97 %
jing	20243	19442	96,04 %	66	16,22 %
zhe	5148	4936	95,88 %	67	16,46 %
zeng	640	612	95,63 %	68	16,71 %
qin	1921	1833	95,42 %	69	16,95 %
geng	342	326	95,32 %	70	17,20 %
zhou	3038	2886	95,00 %	71	17,44 %
zai	3834	3630	94,68 %	72	17,69 %
gu	13825	13081	94,62 %	73	17,94 %
zao	1756	1656	94,31 %	74	18,18 %
deng	2141	2014	94,07 %	75	18,43 %
zou	1350	1266	93,78 %	76	18,67 %
gui	3433	3219	93,77 %	77	18,92 %
ji	*58104*	*54461*	93,73 %	78	19,16 %
mian	1228	1150	93,65 %	79	19,41 %
ruo	298	279	93,62 %	80	19,66 %
xu	3703	3464	93,55 %	81	19,90 %
zui	1999	1869	93,50 %	82	20,15 %
qie	220	205	93,18 %	83	20,39 %
gao	6002	5588	93,10 %	84	20,64 %
luo	1655	1534	92,69 %	85	20,88 %
duan	1726	1597	92,53 %	86	21,13 %
guai	287	265	92,33 %	87	21,38 %
lüe	2361	2179	92,29 %	88	21,62 %
yan	*22830*	*21059*	92,24 %	89	21,87 %
shuo	8103	7472	92,21 %	90	22,11 %
cai	5766	5298	91,88 %	91	22,36 %
jin	10831	9941	91,78 %	92	22,60 %
rong	1646	1510	91,74 %	93	22,85 %
jiong	12	11	91,67 %	94	23,10 %

Syllabe	KW ds Titre	KW ds Titre & la=Chinois	Ratio	Rang	Rang cum.
cong	**46492**	**42528**	**91,47 %**	**95**	**23,34 %**
pian	3104	2827	91,08 %	96	23,59 %
shao	3239	2932	90,52 %	97	23,83 %
wen	*52293*	*47222*	90,30 %	98	24,08 %
neng	1003	905	90,23 %	99	24,32 %
shuang	733	661	90,18 %	100	24,57 %
shen	5403	4867	90,08 %	101	24,82 %
bao	9218	8287	89,90 %	102	25,06 %
dang	8531	7669	89,90 %	103	25,31 %
wai	6134	5496	89,60 %	104	25,55 %
tian	3338	2989	89,54 %	105	25,80 %
ying	11755	10510	89,41 %	106	26,04 %
huan	2795	2487	88,98 %	107	26,29 %
dui	3293	2929	88,95 %	108	26,54 %
dao	6766	6018	88,94 %	109	26,78 %
niao	611	543	88,87 %	110	27,03 %
sheng	22578	20034	88,73 %	111	27,27 %
gai	6197	5494	88,66 %	112	27,52 %
nong	4282	3777	88,21 %	113	27,76 %
lun	24752	21753	87,88 %	114	28,01 %
ge	10263	9018	87,87 %	115	28,26 %
huo	5045	4432	87,85 %	116	28,50 %
tong	11263	9836	87,33 %	117	28,75 %
liang	4643	4046	87,14 %	118	28,99 %
heng	1142	994	87,04 %	119	29,24 %
ming	18547	16133	86,98 %	120	29,48 %
quan	8911	7750	86,97 %	121	29,73 %
shou	7861	6834	86,94 %	122	29,98 %
miao	1290	1118	86,67 %	123	30,22 %
shang	9529	8246	86,54 %	124	30,47 %
zhi	*32643*	*28243*	86,52 %	125	30,71 %
wu	21817	18872	86,50 %	126	30,96 %
liao	11756	10154	86,37 %	127	31,20 %
liu	6748	5818	86,22 %	128	31,45 %
fang	13931	11988	86,05 %	129	31,70 %
kuai	1196	1017	85,03 %	130	31,94 %
fen	6635	5635	84,93 %	131	32,19 %
feng	7385	6261	84,78 %	132	32,43 %
xie	3554	3010	84,69 %	133	32,68 %
hou	2681	2267	84,56 %	134	32,92 %
tiao	2919	2452	84,00 %	135	33,17 %
fa	29413	24669	83,87 %	136	33,42 %
yuan	19129	16035	83,83 %	137	33,66 %
wei	11937	9999	83,76 %	138	33,91 %
ci	10286	8602	83,63 %	139	34,15 %
gua	257	214	83,27 %	140	34,40 %
lu / lü	22646	18845	83,22 %	141	34,64 %
shui	6199	5131	82,77 %	142	34,89 %
ping	7958	6540	82,18 %	143	35,14 %
pin	6128	5025	82,00 %	144	35,38 %
shu	*73897*	*60526*	81,91 %	145	35,63 %
meng	3143	2567	81,67 %	146	35,87 %
yao	7908	6454	81,61 %	147	36,12 %
hui	26200	21378	81,60 %	148	36,36 %
ling	4601	3751	81,53 %	149	36,61 %

Syllabe	KW ds Titre	KW ds Titre & la=Chinois	Ratio	Rang	Rang cum.
hao	4042	3294	81,49 %	150	36,86 %
leng	345	281	81,45 %	151	37,10 %
yin	8725	7100	81,38 %	152	37,35 %
ju	6789	5523	81,35 %	153	37,59 %
bie	658	535	81,31 %	154	37,84 %
peng	875	711	81,26 %	155	38,08 %
shi	*79522*	*64143*	80,66 %	156	38,33 %
chuang	4047	3254	80,41 %	157	38,57 %
hua	34081	27293	80,08 %	158	38,82 %
huang	3237	2582	79,77 %	159	39,07 %
cheng	18104	14413	79,61 %	160	39,31 %
weng	156	124	79,49 %	161	39,56 %
zang	1026	815	79,43 %	162	39,80 %
nuan	102	81	79,41 %	163	40,05 %
kua	598	471	78,76 %	164	40,29 %
lin	5020	3942	78,53 %	165	40,54 %
fei	2843	2225	78,26 %	166	40,79 %
yun	6522	5084	77,95 %	167	41,03 %
jun	4158	3218	77,39 %	168	41,28 %
huai	888	685	77,14 %	169	41,52 %
shan	6545	5045	77,08 %	170	41,77 %
kou	2429	1869	76,95 %	171	42,01 %
bing	2750	2111	76,76 %	172	42,26 %
kuang	4479	3438	76,76 %	173	42,51 %
niang	228	175	76,75 %	174	42,75 %
wang	5570	4275	76,75 %	175	43,00 %
zuan	323	246	76,16 %	176	43,24 %
yong	8308	6313	75,99 %	177	43,49 %
chuan	12683	9529	75,13 %	178	43,73 %
piao	2024	1504	74,31 %	179	43,98 %
dong	7703	5703	74,04 %	180	44,23 %
dou	1142	844	73,91 %	181	44,47 %
rou	133	98	73,68 %	182	44,72 %
xi	24480	17920	73,20 %	183	44,96 %
suan	1754	1283	73,15 %	184	45,21 %
ruan	268	195	72,76 %	185	45,45 %
ding	2477	1801	72,71 %	186	45,70 %
ting	3742	2720	72,69 %	187	45,95 %
pei	5996	4357	72,67 %	188	46,19 %
bai	6977	5066	72,61 %	189	46,44 %
tan	10791	7813	72,40 %	190	46,68 %
chang	11158	8058	72,22 %	191	46,93 %
mei	8219	5935	72,21 %	192	47,17 %
ye	16051	11540	71,90 %	193	47,42 %
tuan	1992	1425	71,54 %	194	47,67 %
fu	12580	8997	71,52 %	195	47,91 %
nan	8144	5821	71,48 %	196	48,16 %
cao	2057	1467	71,32 %	197	48,40 %
kao	8647	6139	71,00 %	198	48,65 %
wan	6514	4611	70,79 %	199	48,89 %
tao	6874	4782	69,57 %	200	49,14 %
tuo	862	596	69,14 %	201	49,39 %
tang	6611	4567	69,08 %	202	49,63 %
yi	*37223*	*25697*	69,04 %	203	49,88 %
gei	413	285	69,01 %	204	50,12 %

Syllabe	KW ds Titre	KW ds Titre & la=Chinois	Ratio	Rang	Rang cum.
bu	7879	5431	68,93 %	205	50,37 %
pao	5318	3649	68,62 %	206	50,61 %
shua	214	146	68,22 %	207	50,86 %
chan	10734	7323	68,22 %	208	51,11 %
mao	5589	3770	67,45 %	209	51,35 %
gan	2663	1794	67,37 %	210	51,60 %
chao	4929	3316	67,28 %	211	51,84 %
chu	23698	15920	67,18 %	212	52,09 %
qia	6	4	66,67 %	213	52,33 %
ning	923	609	65,98 %	214	52,58 %
wo	4919	3241	65,89 %	215	52,83 %
fo	2392	1569	65,59 %	216	53,07 %
mu	6882	4513	65,58 %	217	53,32 %
lie	7757	5061	65,24 %	218	53,56 %
fan	6068	3959	65,24 %	219	53,81 %
lai	3193	2082	65,21 %	220	54,05 %
teng	968	631	65,19 %	221	54,30 %
pu	9308	6046	64,95 %	222	54,55 %
chun	5513	3527	63,98 %	223	54,79 %
seng	238	152	63,87 %	224	55,04 %
cui	1694	1076	63,52 %	225	55,28 %
chou	3426	2170	63,34 %	226	55,53 %
miu	64	40	62,50 %	227	55,77 %
yang	11979	7469	62,35 %	228	56,02 %
shun	329	205	62,31 %	229	56,27 %
tai	15667	9721	62,05 %	230	56,51 %
li	*51995*	*32113*	61,76 %	231	56,76 %
hu	6725	4130	61,41 %	232	57,00 %
kuan	4700	2885	61,38 %	233	57,25 %
beng	112	68	60,71 %	234	57,49 %
ku	14111	8544	60,55 %	235	57,74 %
lao	3932	2369	60,25 %	236	57,99 %
ri	5130	3075	59,94 %	237	58,23 %
shuan	27	16	59,26 %	238	58,48 %
suo	5353	3172	59,26 %	239	58,72 %
hei	974	571	58,62 %	240	58,97 %
keng	287	166	57,84 %	241	59,21 %
chen	7651	4425	57,84 %	242	59,46 %
kang	4781	2746	57,44 %	243	59,71 %
hai	10913	6240	57,18 %	244	59,95 %
ban	6056	3453	57,02 %	245	60,20 %
chi	34256	19459	56,80 %	246	60,44 %
she	25554	14381	56,28 %	247	60,69 %
niu	574	323	56,27 %	248	60,93 %
ti	34524	19219	55,67 %	249	61,18 %
ce	11849	6568	55,43 %	250	61,43 %
tui	2620	1434	54,73 %	251	61,67 %
han	13449	7203	53,56 %	252	61,92 %
kuo	20033	10704	53,43 %	253	62,16 %
er	8401	4338	51,64 %	254	62,41 %
hun	2702	1392	51,52 %	255	62,65 %
cha	7637	3932	51,49 %	256	62,90 %
chai	1652	834	50,48 %	257	63,14 %
sha	1701	855	50,26 %	258	63,39 %
tu	25519	12746	49,95 %	259	63,64 %

Syllabe	KW ds Titre	KW ds Titre & la=Chinois	Ratio	Rang	Rang cum.
tie	2010	989	49,20 %	260	63,88 %
cuan	49	24	48,98 %	261	64,13 %
mou	899	439	48,83 %	262	64,37 %
pang	659	313	47,50 %	263	64,62 %
min	33348	15465	46,37 %	264	64,86 %
gang	4197	1925	45,87 %	265	65,11 %
chui	652	298	45,71 %	266	65,36 %
kan	24524	11133	45,40 %	267	65,60 %
lan	5673	2557	45,07 %	268	65,85 %
nu / nü	6406	2873	44,85 %	269	66,09 %
ran	3172	1393	43,92 %	270	66,34 %
sui	4638	2026	43,68 %	271	66,58 %
bo	5371	2304	42,90 %	272	66,83 %
ben	11703	4864	41,56 %	273	67,08 %
pai	5742	2364	41,17 %	274	67,32 %
hong	6909	2833	41,00 %	275	67,57 %
hang	3140	1263	40,22 %	276	67,81 %
lei	8362	3306	39,54 %	277	68,06 %
dai	*52127*	19886	38,15 %	278	68,30 %
pen	6497	2475	38,09 %	279	68,55 %
zun	307	115	37,46 %	280	68,80 %
luan	866	324	37,41 %	281	69,04 %
ke	24843	8927	35,93 %	282	69,29 %
pi	7575	2620	34,59 %	283	69,53 %
lou	3042	1049	34,48 %	284	69,78 %
kui	600	204	34,00 %	285	70,02 %
mo	10688	3600	33,68 %	286	70,27 %
chong	3774	1269	33,62 %	287	70,52 %
ta	16217	5366	33,09 %	288	70,76 %
kong	6017	1968	32,71 %	289	71,01 %
pou	712	232	32,58 %	290	71,25 %
ze	3980	1283	32,24 %	291	71,50 %
nei	6102	1946	31,89 %	292	71,74 %
si	33425	10310	30,85 %	293	71,99 %
nuo	189	58	30,69 %	294	72,24 %
he	28720	8670	30,19 %	295	72,48 %
mang	650	194	29,85 %	296	72,73 %
pan	10456	3120	29,84 %	297	72,97 %
zan	336	98	29,17 %	298	73,22 %
juan	16134	4596	28,49 %	299	73,46 %
rang	941	264	28,06 %	300	73,71 %
ai	7555	2057	27,23 %	301	73,96 %
sai	1489	401	26,93 %	302	74,20 %
ya	9319	2493	26,75 %	303	74,45 %
sou	527	139	26,38 %	304	74,69 %
kun	2250	588	26,13 %	305	74,94 %
lang	3209	770	24,00 %	306	75,18 %
nao	1966	468	23,80 %	307	75,43 %
di	*189115*	*44360*	23,46 %	308	75,68 %
kai	17399	3917	22,51 %	309	75,92 %
tou	9442	2115	22,40 %	310	76,17 %
duo	3250	703	21,63 %	311	76,41 %
zen	3343	711	21,27 %	312	76,66 %
bi	18820	3971	21,10 %	313	76,90 %
che	12787	2522	19,72 %	314	77,15 %

Syllabe	KW ds Titre	KW ds Titre & la=Chinois	Ratio	Rang	Rang cum.
tun	864	168	19,44 %	315	77,40 %
bang	1696	325	19,16 %	316	77,64 %
san	46321	7726	16,68 %	317	77,89 %
da	*86674*	14299	16,50 %	318	78,13 %
su	34974	5598	16,01 %	319	78,38 %
ba	14048	2222	15,82 %	320	78,62 %
reng	20	3	15,00 %	321	78,87 %
kei	368	53	14,40 %	322	79,12 %
gen	3208	453	14,12 %	323	79,36 %
rao	539	71	13,17 %	324	79,61 %
mi	23428	2994	12,78 %	325	79,85 %
ma	14638	1828	12,49 %	326	80,10 %
bei	28728	3541	12,33 %	327	80,34 %
zei	171	21	12,28 %	328	80,59 %
te	20169	2472	12,26 %	329	80,84 %
nang	668	79	11,83 %	330	81,08 %
zu	*50782*	5866	11,55 %	331	81,33 %
cen	268	30	11,19 %	332	81,57 %
rui	606	66	10,89 %	333	81,82 %
ao	5951	633	10,64 %	334	82,06 %
chua	281	29	10,32 %	335	82,31 %
bin	2099	209	9,96 %	336	82,56 %
mie	728	72	9,89 %	337	82,80 %
sen	3521	324	9,20 %	338	83,05 %
long	9976	901	9,03 %	339	83,29 %
nin	896	74	8,26 %	340	83,54 %
sang	3607	297	8,23 %	341	83,78 %
nüe	112	9	8,04 %	342	84,03 %
nie	1421	112	7,88 %	343	84,28 %
men	30926	2308	7,46 %	344	84,52 %
se	20116	1445	7,18 %	345	84,77 %
you	*99761*	7131	7,15 %	346	85,01 %
fou	596	41	6,88 %	347	85,26 %
shai	231	15	6,49 %	348	85,50 %
re	11131	715	6,42 %	349	85,75 %
sun	14963	862	5,76 %	350	86,00 %
song	45384	2609	5,75 %	351	86,24 %
du	*116542*	6243	5,36 %	352	86,49 %
ang	676	34	5,03 %	353	86,73 %
ru	36328	1708	4,70 %	354	86,98 %
can	28152	1318	4,68 %	355	87,22 %
dan	31568	1323	4,19 %	356	87,47 %
lo	14878	595	4,00 %	357	87,71 %
za	32232	1256	3,90 %	358	87,96 %
mai	16007	597	3,73 %	359	88,21 %
pa	23786	885	3,72 %	360	88,45 %
shei	82	3	3,66 %	361	88,70 %

Syllabe	KW ds Titre	KW ds Titre & la=Chinois	Ratio	Rang	Rang cum.
ni	37000	1322	3,57 %	362	88,94 %
po	51397	1746	3,40 %	363	89,19 %
chuai	30	1	3,33 %	364	89,43 %
ou	21528	703	3,27 %	365	89,68 %
hen	9812	293	2,99 %	366	89,93 %
cou	74	2	2,70 %	367	90,17 %
diu	262	7	2,67 %	368	90,42 %
sao	7108	188	2,64 %	369	90,66 %
cu	3003	70	2,33 %	370	90,91 %
man	*77521*	1753	2,26 %	371	91,15 %
chuo	1244	28	2,25 %	372	91,40 %
an	*182000*	4075	2,24 %	373	91,65 %
dun	14597	321	2,20 %	374	91,89 %
ken	10877	214	1,97 %	375	92,14 %
nai	5719	110	1,92 %	376	92,38 %
gun	5569	80	1,44 %	377	92,63 %
sa	15790	225	1,42 %	378	92,87 %
nia	71	1	1,41 %	379	93,12 %
lia	286	4	1,40 %	380	93,37 %
ng	895	11	1,23 %	381	93,61 %
de	*158138*	1868	1,18 %	382	93,86 %
yo	5723	66	1,15 %	383	94,10 %
run	5993	66	1,10 %	384	94,35 %
pie	1994	14	0,70 %	385	94,59 %
que	32421	223	0,69 %	386	94,84 %
ka	19372	115	0,59 %	387	95,09 %
e / ê	*168550*	891	0,53 %	388	95,33 %
ei	2400	12	0,50 %	389	95,58 %
nou	420	2	0,48 %	390	95,82 %
na	*114054*	525	0,46 %	391	96,07 %
wa	*61170*	244	0,40 %	392	96,31 %
le	*185788*	626	0,34 %	393	96,56 %
ga	6825	22	0,32 %	394	96,81 %
a	*230000*	709	0,31 %	395	97,05 %
eng	5655	17	0,30 %	396	97,30 %
ca	9377	26	0,28 %	397	97,54 %
ha	*53381*	117	0,22 %	398	97,79 %
nen	10678	22	0,21 %	399	98,03 %
die	*174739*	252	0,14 %	400	98,28 %
la	*197718*	203	0,10 %	401	98,53 %
en	*177721*	163	0,09 %	402	98,77 %
ne	10714	9	0,08 %	403	99,02 %
me	74492	56	0,08 %	404	99,26 %
den	*82397*	18	0,02 %	405	99,51 %
dia	5015	1	0,02 %	406	99,75 %
dei	22737	1	0,00 %	407	100,00 %

1. 在过去35年里，日本皇室的9名新生儿全是女的，严重缺乏男性继位人选。日本国会被迫考虑修改宪法，以接纳一位女天皇。
2. Zai guo qu 35 nian li, Ri ben huang shi de 9 ming xin sheng er quan shi nü de, yan zhong que fa nan xing ji wei ren xuan. Ri ben guo hui bei po kao lü xiu gai xian fa, yi jie na yi wei nü tian huang.
3. Zai guoqu 35 nian li, Riben huangshi de 9 ming xinsheng'er quan shi nü de, yanzhong quefa nanxing jiwei renxuan. Riben guohui beipo kaolü xiugai xianfa, yi jiena yi wei nü tianhuang.
4. À noter que ce produit utilise la technologie développée au RALI (Recherche appliquée en linguistique informatique, Université de Montréal) ; il n'est donc pas étonnant d'obtenir les mêmes résultats.
5. Les valeurs en italiques sont estimées.