Meta
Journal des traducteurs
Translators' Journal

META

DICO : un outil de consultation de dictionnaire en réseau

Gilbert Robert and Dominique Petitpierre

Volume 42, Number 2, juin 1997

Lexicologie et terminologie II (1) et Traduction et post-colonialisme en Inde
Translation and Postcolonialism: India (2)

URI: https://id.erudit.org/iderudit/003954ar
DOI: https://doi.org/10.7202/003954ar

See table of contents

Publisher(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (print)
1492-1421 (digital)

Explore this journal

Cite this article

Robert, G. & Petitpierre, D. (1997). DICO : un outil de consultation de dictionnaire en réseau. *Meta*, *42*(2), 283–290. https://doi.org/10.7202/003954ar

Article abstract

DICO is a tool for consulting dictionaries through a computer network. Différent types of access keys permit research to be carried out. Thanks to a modular conception, new dictionaries and specialized indexes can be easily added. Reliable and versatile, DICO offers a wide range of interesting possibilities for growth and development.

DICO : UN OUTIL DE CONSULTATION DE DICTIONNAIRE EN RÉSEAU*

GILBERT ROBERT ET DOMINIQUE PETITPIERRE
ISSCO, Genève, Suisse

Résumé

DICO est un outil de consultation de dictionnaires à travers un réseau informatique. Il offre la possibilité d'effectuer des recherches à l'aide de différents types de clés d'accès. L'addition de nouveaux dictionnaires ou d'index spécialisés est facilitée par une conception modulaire. DICO est fiable, flexible, et offre de nombreuses possibilités d'évolution intéressantes.

Abstract

DICO is a tool for consulting dictionaries through a computer network. Different types of access keys permit research to be carried out. Thanks to a modular conception, new dictionaries and specialized indexes can be easily added. Reliable and versatile, DICO offers a wide range of interesting possibilities for growth and development.

1. INTRODUCTION

L'adoption des ordinateurs et des réseaux informatiques a bouleversé la création, le stockage et la recherche de documents. Cela signifie que la vue classique des dictionnaires imprimés, bien que toujours très utile, devrait être complétée ou remplacée par une nouvelle qui prenne pleinement avantage de la puissance de l'ordinateur. La structure classique d'un mot clé, plus une définition, indexé par ordre alphabétique, bien qu'efficace sur le papier, n'est plus acceptable avec les moyens informatiques actuels. C'est pour combler ce vide que le projet DICO a commencé.

La recherche sur les dictionnaires informatiques a été lancée par l'Université de Waterloo, avec les travaux sur l'*Oxford English Dictionary* (OED) (Weiner 1989). Ce projet a permis la publication de la version électronique de l'OED sur CD-ROM. Plusieurs techniques ont été développées ou adaptées pour atteindre ce but: formateurs interactifs (Darrell 1990), index pour suivre les requêtes contextuelles, utilisation du format SGML (*Standard Generalized Markup Languages*) (Bryan 1988) pour les entrées lexicales (Fawcett 1988), et les puissantes techniques de «string matching» (Baeza-Yates et Gonnet 1992).

Bien que le CD-ROM de l'OED soit livré à un large public, le projet ACQUILEX (Copestake *et al.* 1993), un projet de recherche Esprit, a développé des outils pour l'exploitation des dictionnaires électroniques par les lexicographes (Carroll 1993).

Plus récemment, des chercheurs ont étudié les possibilités offertes par l'ouverture des connexions informatiques à travers Internet. Ils ont montré le bénéfice des interfaces uniformes pour l'accès à l'information hétérogène ainsi que les besoins pour des ressources partagées (Rao *et al.* 1995).

DICO est un outil de consultation de données lexicales utilisant le réseau informatique Internet. Il a été conçu pour être utilisé par des traducteurs et des utilisateurs non spécialistes. L'originalité et la puissance de cet outil réside dans sa capacité à être indépendant du format des données et dans un ensemble de fonctions de recherche de base. Il permet aussi d'ajouter aisément des index spécialisés (pour lexicographes ou

linguistes). De plus, les entrées lexicales peuvent être vues dans des formats d'affichages complètement redéfinissables.

Son architecture client-serveur offre plusieurs possibilités : d'abord une seule copie des données est nécessaire sur le réseau, et ensuite, l'interface client peut être réécrite pour un environnement informatique spécifique. Actuellement, deux interfaces sont disponibles : **xdico** pour les utilisateurs sous X-Windows et **tdico** pour les utilisateurs d'un terminal plein écran. **DICO** est implanté sous Unix.

Les avantages de **DICO** sur ces produits commerciaux largement répandus se résument en six points :

- **interface unique** quel que soit le dictionnaire (éditeur, langue, etc.) ;
- **indépendant du format** de représentation des données ;
- **architecture client-serveur** : une seule copie des données sur le réseau est nécessaire ;
- **fonctions de recherche** : extensibles ;
- **affichage** configurable suivant les besoins ;
- **sécurité** : transmission sophistiquée sur le réseau et protection des données par cryptage ;
- **disponibilité** : facilitée pour des dictionnaires spécialisés ou conçus par des éditeurs sans grands moyens financiers.

L'outil **DICO** est opérationnel sur le réseau de l'Université de Genève et permet de consulter 10 dictionnaires monolingues et bilingues. **DICO** a été développé sous le mandat de l'association SUISSETRA et a été présenté au stand Suisse du **CeBIT**'95 à Hanovre comme l'un des 12 gagnants du concours *La Suisse Carrefour des Technologies*. Il a été primé en tant qu'exemple de nouvelles solutions technologiques et en raison de son caractère innovateur.

L'évolution actuelle de **DICO** est axée sur une redéfinition plus large de l'aspect réseau, des protocoles de communication et des interfaces. Les connaissances linguistiques de l'**ISSCO** seront pleinement introduites dans cette optique, notamment les outils morphologiques (Bouillon 1995) (Russell *et al.* 1992) et les outils de décodage et de conversion de document (Armstrong-Warwick 1993, 1994).

Dans les pages qui suivent, nous présenterons tout d'abord l'architecture du système et la description de ses différentes parties ; nous envisagerons ensuite les différentes fonctions de recherche et d'affichage, et enfin les possibilités d'évolution que le groupe de l'ISSCO envisage pour DICO.

2. ARCHITECTURE

Le système est divisé en trois parties distinctes :

- un programme serveur de dictionnaires qui gère les requêtes émises par les utilisateurs ;
- un programme serveur d'informations (DIS : *Dictionary Information Service*), chargé d'indiquer au programme client la liste des ressources disponibles avec les différents sites où ils sont présents ;
- un programme client chargé de dialoguer avec l'utilisateur et les serveurs.

Le serveur d'informations est la tour de contrôle du système. Sa fonction première est de contrôler l'appartenance du client qui désire se connecter au groupe des utilisateurs autorisés. Par la suite, il informera le client des divers emplacements des serveurs de données (adresse Internet). Cette centralisation permet de gérer aisément l'installation (ou la suppression) de nouveaux serveurs ou de nouveaux dictionnaires.

Le support physique de cette architecture est le réseau Internet et son protocole de communication TCP/IP (*Transfert Control Protocol/Internet Protocol*). Son essor actuel, avec l'explosion des serveurs Web et la publicité engendrée, nous laisse penser que nous avons fait le bon choix.

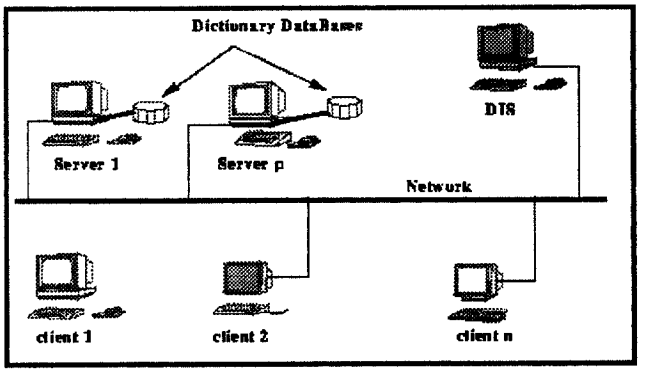

Figure 1 :
Architecture de DICO

Le serveur de données

Un serveur de données peut servir plusieurs dictionnaires et divers clients. Il récupère les requêtes des utilisateurs à travers le réseau, accomplit l'opération associée à cette requête et donne la réponse. La principale opération consiste à trouver quelle entrée du dictionnaire correspond à la clé de recherche donnée et à retrouver le contenu de cette entrée. Des opérations standard sont disponibles telles que :

- afficher la prochaine (précédente) entrée dans la séquence du dictionnaire ;
- afficher les informations associées (liste des abréviations, auteurs, copyright) ;
- changer le mode de recherche ainsi que le format d'affichage.

Chaque serveur de dictionnaire sur le réseau peut servir plusieurs requêtes émanant de clients différents.

Pour DICO, un dictionnaire est une collection d'entrées auxquelles est associé un mot clé. Ces mots clés constituent les clés d'accès primaires ; elles peuvent être retrouvées par correspondance exacte, par préfixe ou par expressions régulières (semblables à la commande *grep* d'UNIX). Il est aussi possible de faire une recherche en utilisant les données incluses dans l'entrée lexicale.

Il s'agit alors d'un mode d'accès secondaire par clé. Par exemple, on peut demander toutes les entrées dans un dictionnaire bilingue dont le champ de traduction contient le mot «décorative».

3. L'INTERFACE UTILISATEUR

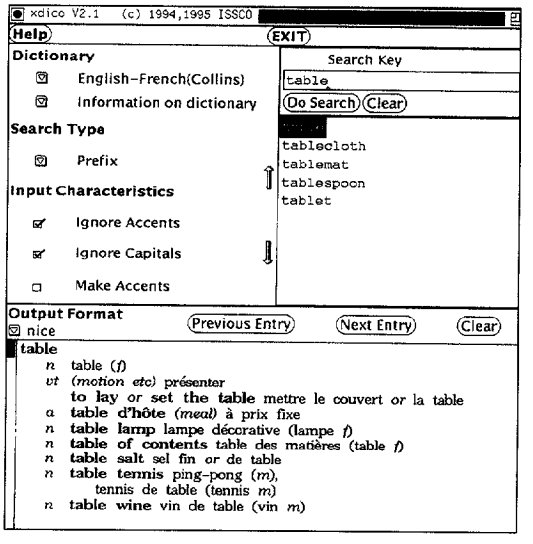

Figure 2 :
Interface pour X-Windows xdico

L'interface utilisateur montre d'abord sur l'écran de l'utilisateur ce qui est disponible ; elle prend ensuite en entrée les changements de configuration (affichage, mode de recherche...) ou les requêtes ; elle communique avec le serveur de dictionnaires à travers le réseau et affiche le résultat.

La réponse à une question composée d'une clé de recherche est une liste d'en-têtes (exemple dans la figure 2 : avec la clé de recherche *table* en mode Prefix, le serveur a donné cinq solutions). Celles-ci sont affichées sur l'écran et l'utilisateur peut choisir de voir une ou plusieurs entrées complètes dans la partie inférieure de l'écran. D'autres interactions sont possibles grâce aux menus ou aux boutons, au déroulement de l'affichage, ainsi qu'au moyen des pages d'aide.

Le mode d'affichage des entrées est complètement redéfinissable à l'aide d'un transducteur (voir la figure 3 : un automate à état fini). Cela permet d'exploiter les possibilités graphiques de la machine que l'on utilise. Le transducteur produit des instructions élémentaires indépendantes de la machine qui seront interprétées par un formateur et un afficheur. Ainsi, pour chaque dictionnaire qui possède un format particulier, un ensemble de formats d'affichages (de grammaire) sera défini. Ces informations sont récupérées par le serveur sur lequel, pour chaque dictionnaire, existent au moins deux formats : *source* (style dans lequel les données sont stockées, SGML) et *nice* (style dans lequel les indentations et les fontes permettent d'afficher au mieux les entrées).

Cette méthode facilite l'introduction d'un nouveau dictionnaire (codé dans un nouveau format), pour lequel il ne faudra écrire qu'une nouvelle grammaire de transduction.

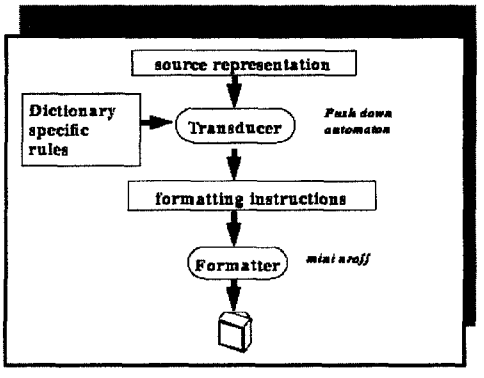

Figure 3 :
Le transducteur

4. LES FONCTIONS SPÉCIALES

Dans sa requête, l'utilisateur a la possibilité de spécifier s'il veut ignorer ou non des informations dans sa requête. Par exemple :

- **ignorer les majuscules** : à la clé de recherche *march* pourra correspondre *March* et *march* ;
- **ignorer les caractères accentués** : à la clé *eleve* pourra correspondre *élève* et *élevé* ;
- **ignorer les ponctuations et espaces** : à la clé *bb* ou *b&b* correspondra *b.&b.* et à *crowsnest* correspondra *crows'nest.*

De plus, comme il n'est pas garanti dans un environnement hétérogène qu'il soit possible de taper et d'afficher les caractères accentués, l'interface utilisateur propose des modes spéciaux comme la possibilité de taper les caractères *a`* pour remplacer le *à* ou *a:* pour *ä.*

Enfin, une attention toute particulière a été apportée à la protection des données, un point sensible pour la plupart des créateurs et fournisseurs de données. Ce point important est essentiel pour l'obtention des dictionnaires et pour se conformer à des droits d'auteurs très stricts. Bien entendu, il n'est pas possible de garantir à cent pour cent la fiabilité d'un système, mais les barrières dressées sont suffisamment étudiées pour limiter au maximum des problèmes de piratage. La photocopie du dictionnaire imprimé associée à un programme de reconnaissance de caractères serait beaucoup plus facile et rapide.

Les données sont encryptées sur les serveurs, ainsi que sur le réseau avec une clé différente pour chaque nouvelle connexion. Le nombre d'accès successifs (*next entry*) au même dictionnaire est sanctionné, et l'accès est réservé uniquement aux ayants droit.

5. REPRÉSENTATION DES DONNÉES ET INDEXATION

Bien que cet outil ne requière pas de format particulier pour les données, la plupart des dictionnaires doivent être convertis d'un format typographique original à un format standard SGML (spécifique à chacun). Cette opération exige une expérience dans le domaine du codage des données, du fait :

- des erreurs de représentation d'ordre typographique, ou de corruption de fichier ;
- de l'inconsistance des entrées lexicales (différents lexicographes à différentes périodes) ;
- des marquages souvent absents ou incompréhensibles.

Dans ce but, nous avons suivi les lignes directrices du TEI (*Text Encoding Initiative*), et nous avons ainsi rendu les données beaucoup plus lisibles.

Pour DICO, ce traitement était un pas nécessaire pour identifier les structures des entrées lexicales et indexer les parties adéquates. Ainsi, nous avons construit différents types d'index :

- mots apparaissant dans la définition ou la traduction de l'entrée ;
- recherche sur la forme canonique des mots, faisant intervenir un traitement morphologique. Si nous cherchons le mot *courir* dans le champ traduction du dictionnaire anglais-français, le système trouvera les entrées :

  accrue : dans laquelle apparaît : **accrue interest : *intérêt couru***
  pu : dans laquelle apparaît : **vt (rumour) *faire courir***

- recherche des domaines d'information associés à la définition ou à la traduction (exemple : *BIOLOGY, LAW, THEATRE...*) ;
- recherche des anagrammes ;
- recherche des entrées suivant le mode *Scrabble*.

Un point très important qui est apparu au cours de ce projet concerne l'ordre alphabétique. Ce problème apparaît lorsqu'on observe une suite d'entrées solutions ou que l'on demande une entrée suivante. Les règles de tri sont différentes suivant les langues.

- En danois ***æ*** est trié après la lettre ***z***.
- En espagnol le ***ch*** est trié après ***cz*** et le ***ll*** après ***lz***.
- En français l'accent sur un caractère est utilisé comme une clé secondaire. Par exemple, les mots suivants seront ordonnés alphabétiquement comme :

  mais
  maïs
  maison

Dans **DICO** une description de l'ordre alphabétique est associée à chaque dictionnaire.

6. IMPLÉMENTATION

DICO est programmé en C et utilise les librairies standard. Il fonctionne sous le système d'exploitation d'UNIX (SunOs 4.1.x et Solaris 2.x et X-Windows pour xdico). Une version est implantée à l'Université de Genève avec la mise à disposition de 10 dictionnaires (avec copyright) :

- six dictionnaires bilingues (format poche) de Harper-Collins (français-anglais, français-allemand, anglais-français, allemand-français, anglais-allemand et allemand-anglais) ;
- les deux volumes de l'*Oxford Dictionary of Current Idiomatic English* (verbes et phrases) ;
- le *Oxford Advanced Learner's Dictionary of Current English* ;
- et, dernièrement, le français-suisse-allemand.

Les utilisateurs non spécialistes s'aventurent rarement dans des modes de recherche particuliers et restent le plus souvent dans les modes par défaut (type de recherche par préfixe et type d'affichage *nice*). Ils apprécient tout particulièrement la rapidité de la connexion, et le fait de pouvoir consulter plusieurs dictionnaires sans changer d'interface ou de CD-ROM.

Les utilisateurs spécialistes sont, quant à eux, plus critiques, mais apprécient la démarche et la puissance du système tout en suggérant des fonctionnalités plus complexes, ce qui nous a amené à redéfinir **DICO** dans un projet en cours de réalisation.

7. CONCLUSIONS

Nous avons décrit ici l'outil DICO mis au point à l'ISSCO (Université de Genève). Cet outil marque une nouvelle génération dans la consultation de données lexicales. L'avènement d'Internet montre à quel point le besoin de nouvelles technologies de communication est fort. En effet, en l'espace d'un an, les notions d'hypertexte et de multimédia font partie de la plupart des nouveaux programmes de recherche. On ne compte plus les sites World-Wide-Web mettant à disposition des dictionnaires, mais souvent ils doivent faire face aux droits d'auteur, largement ignorés, de plus les données ne sont pas toujours intéressantes, la communication est lente et l'interface peu adaptée aux caractéristiques spécifiques des données.

Nous avons pu constater, avec l'Université de Genève, la fiabilité du système, et récolter un ensemble très intéressant de commentaires sur la convivialité, l'efficacité et les améliorations à apporter à l'outil. De plus, nous avons eu la chance de participer au CeBIT'95 et d'y exposer nos travaux. Cela nous a permis de mesurer l'intérêt :

- des universités, qui possèdent souvent beaucoup de ressources sans pouvoir les consulter aisément ;
- des particuliers, entre autres les interprètes et les traducteurs, qui possèdent leurs propres bases de données rudimentaires, et qui doivent se déplacer dans les bibliothèques ou se connecter à des services chers et peu efficaces ;
- des entreprises, dont le partage des ressources se limite aux bases de données relationnelles.

Devant ce constat, nous avons décidé de redéfinir l'outil :

- en utilisant plus largement nos connaissances linguistiques ;
- en étendant la notion de réseau, ce qui est déjà le cas actuellement, mais en introduisant les possibilités de réseaux locaux et de réseaux plus larges ;
- en tirant partie de la puissance de SGML ;
- en raffinant l'interface utilisateur ;
- en augmentant les possibilités de recherche dans plusieurs dictionnaires avec peut-être des combinaisons logiques.

DICO est un outil qui a de l'avenir, et un groupe de l'ISSCO travaille dans ce sens. Si les objectifs semblent clairs du point de vue technologique, le plus dur reste toujours les négociations avec les éditeurs qui ont franchi un grand pas avec la distribution à grande échelle des CD-ROM mais qui restent toujours un peu réfractaires aux nouveautés.

Remerciements

Nous tenons à remercier Pierrette Bouillon pour sa patience et la relecture de cet article ainsi que Susan Armstrong et Afzal Ballim pour leurs suggestions.

Note

* Cet article est issu d'une communication présentée par l'auteur aux IV[es] Journées scientifiques du réseau «Lexicologie, terminologie, traduction» de l'AUPELF-UREF (Lyon, France, 28, 29, 30 septembre 1995).

RÉFÉRENCES

Advanced Learner's Dictionary of Current English (1989) : Third Edition, Electronic edition coded with SGML, Oxford University Press.

ARMSTRONG-WARWICK, S. (1993) : «Acquisition and Exploitation of Textual Resources for NLP», *Proceedings of Knowledge Base and Knowledge Systems*, Tokyo, 1993.

ARMSTRONG-WARWICK, S., THOMPSON, H., MCKELVIE, D. et D. PETITPIERRE (1994) : «Data in Your Language: The ECI Multilingual Corpus 1», *Proceedsings of the International Workshop on Sharable Natural Language Resources*, Nara, pp. 97-106.

BAEZA-YATES, R. and G. GONNET (1992) : «A New Approach to Text Searching», *Communication of the ACM*, 35 (10), October, pp. 74-82.

BENMRAD, M. (1994) : *Agrégats et composition de requêtes dans les hypertextes virtuels*, Thèse de doctorat n° 1284, Département d'informatique, École Polytechnique Fédérale de Lausanne, octobre 1994.

BENMRAD, M. CORAY, G. et C. VANOIRBEEK (1995) : «Designing Virtual Hypertexts with Aggregates», *Proceedings of IWHD'95*, Montpellier, juin 1995.

BRYAN, M. (1988) : *SGML An Author's Guide to the Standard Generalized Markup Language*, Wokingham, Addison-Wesley, 5.23.

CARROL, J. (1993) : *Lexical Database System, User Manual*, ESPRIT BRA 3030 Computer Laboratory, University of Cambridge, United Kingdom.

COPESTAKE, A, SANFILIPPO, A., BRISCOE, T. and V. DE PAIVA (1993) : «The Acquilex LKB: An Introduction», Briscoe, De Paiva, Copestake (Eds), *Inheritance, Defaults and the Lexicon*, CUP.

DARRELL, R. Raymond (1990) : «An interactive Formatter for Tagged text», *OED-90-02*, Centre for the New Oxford English Dictionary and Text Research University of Waterloo, Waterloo, Ontario, Canada.

FAWCETT, H. (1988) : *The User's Guide to Pat Centre for the New OED*, University of Waterloo.

GORCY (1989) : «La constitution de la documentation du *Trésor de la langue française* : problèmes et méthodes, *Proceedings of the Fifth Annual Conference of the UW Centre for the New Oxford English Dictionary*, Oxford, p. 33.

HEYLEN, A., MAXWELL, A. K. and S. WARWICK-ARMSTRONG (1989) : «Collocations, Dictionaries and Machine Translation», *Proceedings of the AAAI Symposium on Machine Translation and the Lexicon*, Stanford (Calif.), USA.

IDE, N., VERONIS, J., WARWICK-ARMSTRONG, S. and N. CALZOLARI (1992) : «Principles for Encoding Machine-readable Dictionaries», *EURALEX'92 Proceedings*, Tampere, Finland, pp. 239-246.

Oxford English Dictionary on Compact Disc (1992) : Oxford University Press, Oxford.

PETITPIERRE, P. et G. ROBERT (1995) : «DICO, Technologiestandort Schweiz», *CeBIT 1995*, Hannover.

PETITPIERRE, P., ROBERT, G. et S. ARMSTRONG (1994) : «Design of an On-line Dictionary Consultation Tool».

RAO, R., PEDERSEN, J. O., HEARST, M. A., MACKINLAY, J. D., CARD, S. K., MASINTER, L., HALVORSEN, P.-K. and G. G. ROBERTSON (1995) : «Rich Interaction in the Digital Library Communication», *ACM*, 38 (4), April 1995, pp. 29-39.

RUSSELL, G., BALLIM, A., CARROLL, J. and S. WARWICK-ARMSTRONG (1992) : «A Practical Approach to Multiple Default for Unification-based Lexicons», *Computational Linguistics*, 18 (3).

SPERBERG-MCQUEEN, C. M. and L. BURNARD (Eds) (1994) : *Guidelines for Electronic Text Encoding and Interchange*.

The Unicode Standard (1990) : Worldwide Character Encoding — Version 1.0, Menlio Park (Calif.), Addison-Wesley.

VANOIRBEEK, C. (1992) : «Formatting Structure Tables Proceedings Electronic Publishing», *Document Manipulation and Typography*, Lausanne, April 1992, pp. 291-310.

WARWICK, S. (1994) : «Automated Lexical resources in Europe: A Survey», D. Walker and A. Zampolli (Eds), *Automating the Lexicon*, Clarendon Press.

WEINER, E. S. C. (1989) : «Editing the OED in Electronic Age», *Proceedings of the Fifth Annual Conference of the UW Centre for The New Oxford English Dictionary*, Oxford, pp. 23.

WU, S. and U. MANBER (1992) : «Fast Text Searching Allowing Errors», *Communications of the ACM*, 35 (10), pp. 83-91.