Meta
Journal des traducteurs
Translators' Journal

META

Base de données textuelles et terminographiques

Maria Teresa Rijo da Fonseca Lino

Volume 39, Number 4, décembre 1994

Hommage à Bernard Quemada : termes et textes

URI: https://id.erudit.org/iderudit/003951ar
DOI: https://doi.org/10.7202/003951ar

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

Rijo da Fonseca Lino, M. (1994). Base de données textuelles et terminographiques. *Meta*, *39*(4), 786–789. https://doi.org/10.7202/003951ar

Article abstract

Textual computer science is a relatively new area which has greatly contributed to the rapid evolution of lexicography and terminology. Since January 1991 a textual data base has been set up in the Universidade nova de Lisboa. This Portuguese textual data base is made up largely of popular scientific and technical texts with some everyday language texts. Using hypertext, specific data processing operations can be carried out, which is an important phase in automating terminology and lexicography work. This article focuses on the scanning phase, and on the criteria used for structuring the terminology data base, such as the different processes for selecting scientific texts for terminological research and the use of hypertexts.

BASE DE DONNÉES TEXTUELLES ET TERMINOGRAPHIQUES

MARIA TERESA RIJO DA FONSECA LINO
Université Nouvelle de Lisbonne, Lisbonne, Portugal

Résumé

L'informatique d'orientation textuelle, domaine relativement nouveau, a beaucoup contribué à une évolution rapide de la lexicographie et de la terminologie.

Depuis janvier 1991, nous organisons une base de données textuelles du portugais dans le cadre du Centre d'Études Comparées de l'Université Nouvelle de Lisbonne. Cette base est constituée surtout par des textes scientifiques et techniques, de vulgarisation, mais aussi par un nombre important de textes de langue courante.

Dans cet article, nous mettons l'accent sur la scannérisation et les critères d'organisation de la base textuelle : la sélection de textes scientifiques concernant les différentes recherches en terminologie ; l'utilisation d'hypertextes.

Abstract

Textual computer science is a relatively new area which has greatly contributed to the rapid evolution of lexicography and terminology.

Since January 1991 a textual data base has been set up in the Universidade nôva de Lisboa. This Portuguese textual data base is made up largely of popular scientific and technical texts with some everyday language texts.

Using hypertext, specific data processing operations can be carried out, which is an important phase in automating terminology and lexicography work.

This article focuses on the scanning phase, and on the criteria used for structuring the terminology data base, such as the different processes for selecting scientific texts for terminological research and the use of hypertexts.

Aux concepts de terminotique et de lexicomatique (Quemada 1987) vient s'ajouter l'informatique d'orientation textuelle qui rend compte d'un nouveau domaine de recherche qui a beaucoup contribué à l'évolution rapide de méthodologies et de modèles sémantiques de description de la lexicographie et de la terminographie.

Parmi les différents moyens mis à la disposition de la terminographie assistée par ordinateur, nous soulignons l'importance des hypertextes. Ce sont ces logiciels qui rendent possibles la gestion et l'automatisation du travail sur les textes électroniques qui, dans un premier temps, ont été informatisés.

Étant donné l'importance de la création des corpus informatisés dans le cadre du Réseau Européen des Corpus Textuels — RECT, les chercheurs qui travavaillent sur la langue portugaise ne pouvaient que suivre ces initiatives.

C'est dans ce contexte que, depuis janvier 1991, nous avons décidé de créer une base textuelle à l'Université Nouvelle de Lisbonne (en coopération avec le CNRS — Université de Nice).

Cette base gérée, en partie, par le logiciel Hyperbase, est constituée par des textes saisis par lecture optique et quelquefois par téléchargement ou par saisie manuelle.

Le lecteur optique est rapide et intelligent — même s'il requiert un certain nombre de corrections. Nous sommes attentifs aux propositions en rapport avec les conversions en formats SGML et aux étiquetages normalisés des corpus textuels (*Text Encoding Initiative* — TEI).

Avant l'informatisation, les textes sont soumis à une sélection rigoureuse et différenciée en fonction de recherches de groupe et/ou individuelles; celles-ci ont généralement comme but des DEA ou des doctorats en terminologie et lexicographie spécialisée.

Ce corpus textuel rassemble plusieurs types de textes portugais:

- textes scientifiques et techniques qui constituent le contenu principal de la base; ce sont des textes rédigés par des spécialistes, pour un public du même niveau: des revues spécialisées, des livres, des extraits de thèses de doctorat soutenues récemment;
- textes de «semi-vulgarisation»: des textes de revues élaborés par des scientifiques, ayant comme objectif une «certaine» vulgarisation d'expériences pour des collègues appartenant aux mêmes groupes professionnels;
- textes de «banalisation» (Galisson 1978) scientifique ou technique, recueillis dans des livres, des revues ou des journaux;
- textes d'introduction à une langue de spécialité ou à une terminologie dans un certain domaine; dans ce cas-là, nous avons informatisé des manuels de biologie, utilisés dans les deux années terminales du secondaire;
- textes de vulgarisation scientifique et technique selectionnés dans la presse;
- textes de langue courante: textes sélectionnés dans la presse, dont la principale caractéristique est le grand nombre de néologismes de langue courante et des néologismes résultant de phénomènes de vulgarisation de «termes»;
- un petit sous-ensemble constitué de textes littéraires et publicitaires.

Parallèlement à ce corpus de textes du portugais du Portugal, nous en avons organisé un autre, de dimension moins importante, qui à l'heure actuelle est constitué de textes du portugais de Guinée-Bissao et du Mozambique; pour ce travail qui se développe actuellement, nous comptons sur la collaboration de collègues universitaires de ces pays.

Cette base textuelle est venue enrichir les procédures d'analyse et de description des recherches qui utilisent souvent en parallèle les matériaux de la «Base de néologismes du portugais contemporain» (Lino 1992), de la «Base de néologismes du français contemporain» (base de dimensions moins importantes) et la «Base de terminologie — PORTERM».

Soulignons alors quelques objectifs de ce corpus textuel en ce qui concerne le travail en terminographie (monolingue, bilingue et/ou trilingue):

- sélectionner des unités terminologiques et des néonymes;
- suivre l'implantation de néonymes et les phénomènes de terminologisation dans différents textes d'une certaine période;
- étudier les aspects conceptuels et linguistiques associés à l'apparition d'une notion ou d'un concept: a) la première dénomination; b) les différents synonymes associés à une notion nouvelle (surtout dans les terminologies relatives aux sciences dites «molles»); c) les définitions souvent peu précises;
- sélectionner plusieurs types de contextes: définitoires, fonctionnels, associatifs ou autres;
- délimiter des définitions stabilisées et/ou harmonisées;
- étudier plusieurs phénomènes relatifs aux «internationalismes terminologiques»: des lexies simples ou complexes, des formants, et autres;
- observer des collocations et des phrasèmes en contraste avec d'autres langues (français, anglais);
- observer des processus de néologie et de terminologisation du portugais dans les pays de langue portugaise avec lesquels nous organisons, depuis 1991, un «Réseau de Néologie et de Terminologie de langue portugaise» (Lino 1992); suivre des phénomènes

de socioterminologie ou d'ethnoterminologie qui rendent compte du dialogue avec les cultures locales ;

■ servir de base à des travaux en lexicographie spécialisée pour un public de spécialistes ;

■ préparer des matériaux en lexicographie spécialisée informatisée d'apprentissage pour des contextes scolaires ou pour des situations d'apprentissage de terminologie en contextes non scolaires (entreprise, et autres) (Lino 1991).

Nous voyons donc que la notion de corpus monolingue informatisé est relativement nouvelle. Cependant, plus récemment, les concepts de bitextualité, de corpus bilingue ou trilingue sont venus enrichir la réflexion théorique et les méthodologies en terminographie souvent associées aux technologies de la «nouvelle lexicographie» (Quemada 1990b).

Ainsi, la constitution d'un corpus textuel implique un certain nombre d'hypothèses sur les caractéristiques des discours, sur les relations entre textes et termes que l'on veut analyser ; mais un corpus implique aussi l'existence de modèle(s) interprétatif(s) que l'analyse permettra de corriger et de compléter.

Les fragments du corpus transformés en textes électroniques présentent des avantages méthodiques en ce qui concerne la représentation lexicale ; la transformation permet au chercheur en terminologie d'effectuer des lectures thématiques de son texte,

> «des lectures instantanées qui sont autant de coupes transversales dans le matérial textuel [...] ; en rendant explicite l'axe paradigmatique [...], le lexique apparaît un peu comme un rayon x qui révèle de quoi le texte est fait» (Daoust 1990 : 58).

Ce travail sur des textes et des occurrences concrètes vient compléter l'information souvent organisée dans les bases de terminologie.

Le corpus textuel et les termes se présentent comme sources de données :

> «Le terme est une unité sémantique fondamentale de la langue savante (système et texte). Outre son importance sur le plan lexico-sémantique, le terme est l'instrument essentiel de la cohérence des textes savants, le porteur des sèmes thématiques et du contenu. Il représente les nœuds du réseau isotopique, reflète le niveau raisonné d'intellectualisation et le degré circonstancié de particularisation du texte» (Kocourek 1991 : 74).

L'importance du corpus textuel est incontestable. Étant donné qu'il n'y a pas de textes scientifiques dans une grande partie des domaines du savoir en langue portugaise, nous avons fait le relevé des termes, des définitions et des contextes auprès des spécialistes ; quelquefois même, nous avons recueilli des termes en situation de communication orale spécialisée. Nous voulons ici référer à la collaboration d'experts qui ont un rôle de dynamisateurs dans leurs universités (Médecine générale ; Sénologie ; Physique ; Télédétection ; Audiovisuel).

RÉFÉRENCES

BAKER, M., FRANCIS, G. et E. TOGNINI-BONELLI (1993) : *Text and Technology*, Philadelphia / Amsterdam, John Benjamins Publishing Company.

BRUNET, Étienne (1991) : «Hyperbase», *CUMFID*, 17, Université de Nice.

CANDEL, Danielle (coordination) (1990) : *Dictionnairique et lexicographie*, Paris, Didier Érudition.

CAETANO-MOCHO, Mª. do Céu (à paraître) : «Base de données textuelles : processus néologiques de la langue courante», *CUMFID*.

CONTENTE, Madalena (à paraître) : «Base de données textuelles et lexicographie informatisée d'apprentissage (domaine d'application : biologie)», *CUMFID*.

COSTA, Mª. Rute (à paraître) : «Base de données textuelles : terminologie de l'économie monétaire», *CUMFID*.

DESMET, Isabel (à paraître) : «Les enjeux linguistiques de l'enseignement du portugais spécialisé : la valeur heuristique du plan textuel», *CUMFID*.

DAOUST, François (1990) : «L'informaticien, le lecteur et le texte. L'approche SATO», *ICO — Gestion de l'information textuelle*, 2-3, Montréal.
GALISSON, Robert (1978) : *Recherche de lexicologie descriptive : la banalisation lexicale*, Paris, Nathan.
GARCIA, Mª. de Lurdes (à paraître) : «Base de données textuelles : corpora trilingue et étude des internationalismes terminologiques de la sénologie», *CUMFID*.
KOCOUREK, Rostislav (1991) : «Textes et termes», *Meta*, 36-1, Montréal, Presses de l'Université de Montréal, pp. 71-76.
LINO, Maria Teresa (1991) : «Um projecto em terminodidáctica», *Actas do Encontro do programa ERCI*, Lisboa, Universidade Aberta.
LINO, Maria Teresa (à paraître) : «Lexicografia e terminologia, Seminário O português, língua de comunicação internacional», Lisboa (Actes en préparation).
LINO, Maria Teresa (à paraître) : «Définition d'une base texuelle», *CUMFID*.
MARTIN, Éveline (coordination) (1992) : *Dictionnairique et lexicographie*, 2, Paris, Didier Érudition.
MARTIN, Éveline (1993) : *Reconnaissance de contextes thématiques dans un corpus textuel. Éléments de lexico-sémantique*, Paris, Didier Érudition.
QUEMADA, Bernard (1987) : «Notes sur lexicographie et dictionnairique», *Cahiers de lexicologie*, 51, p. 233.
QUEMADA, Bernard (1990a) : «Lexicographie», *Lexikon der Romanistischen Linguistik (LRL)*, vol. V, 1, Tübingen, Max Niemeyer.
QUEMADA, Bernard (1990b) : La nouvelle lexicographie, *La Linguistica Aplicada*, Barcelona, pp. 56-57.
SCHAETZEN, Caroline de (1993) : «Un accès rapide aux collocations», *Terminologies nouvelles*, n° 10.
SINCLAIR, J. McH. (1991) : «Corpus, concordance, collocation», Oxford, Oxford University Press.