Meta
Journal des traducteurs
Translators' Journal

META

La lexicographie assistée par ordinateur. L'expérience d'UZEI

Jose Antonio Aduriz et Miriam Urkia

Volume 42, numéro 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/001856ar
DOI : https://doi.org/10.7202/001856ar

Aller au sommaire du numéro

Éditeur(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (imprimé)
1492-1421 (numérique)

Découvrir la revue

Citer cet article

Aduriz, J. A. & Urkia, M. (1997). La lexicographie assistée par ordinateur. L'expérience d'UZEI. *Meta*, *42*(2), 257–263. https://doi.org/10.7202/001856ar

Résumé de l'article

Cet article présente l'expérience conduite à UZEI concernant l'utilisation des divers outils informatiques dans l'activité lexicographique : l'outil RTerm, créé pour le dépouillement systématique de textes et pour la lemmatisation, l'édition des listes et l'obtention de statistiques ; ORACLE, base de données relationnelle installée sur un système ouvert UNIX ; et EUSLEM, un lemmatiseur automatique, en cours de création, basé sur l'analyse morphologique du basque.

LA LEXICOGRAPHIE ASSISTÉE PAR ORDINATEUR. L'EXPÉRIENCE D'UZEI*

JOSE ANTONIO ADURIZ ET MIRIAM URKIA
UZEI, Donostia, Espagne

Résumé

Cet article présente l'expérience conduite à UZEI concernant l'utilisation des divers outils informatiques dans l'activité lexicographique : l'outil RTerm, *créé pour le dépouillement systématique de textes et pour la lemmatisation, l'édition des listes et l'obtention de statistiques ;* ORACLE, *base de données relationnelle installée sur un système ouvert UNIX ; et* EUSLEM, *un lemmatiseur automatique, en cours de création, basé sur l'analyse morphologique du basque.*

Abstract

This article describes UZEI's experience using various computerized tools in lexicographic research. These include Rterm, *created for systematic examination of texts, lemmatization, establishing lists and statistics;* ORACLE, *a relational database in an open UNIX system and* EUSLEM, *a lemmatiser (under completion) based on a morphological analysis of Basque.*

L'UZEI (Centre basque pour la normalisation linguistique) est une association culturelle, créée en 1977, ayant pour but de donner une impulsion à la modernisation du lexique basque dans le processus de normalisation. Trois mots définissent le projet d'UZEI : recherche, normalisation et modernisation de la langue basque, l'euskera.

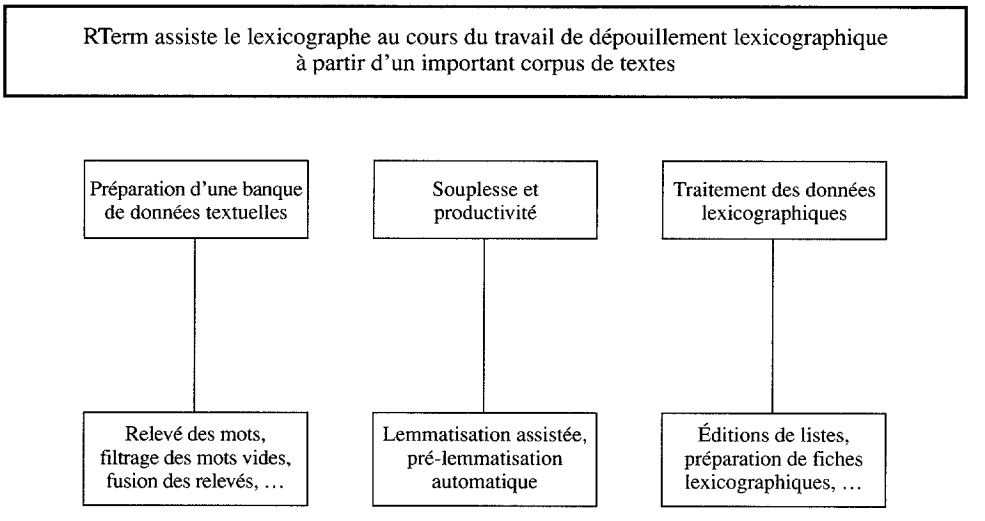
Au cours des premières années, l'association s'est occupée surtout de terminologie, comme en témoignent les 35 dictionnaires terminologiques multilingues déjà publiés. En 1986, cette activité a atteint sa maturité avec la création d'EUSKALTERM (Centre Basque de Terminologie). EUSKALTERM est aussi une banque de données terminologiques qui peut être consultée par videotex et qui est constamment mise à jour.

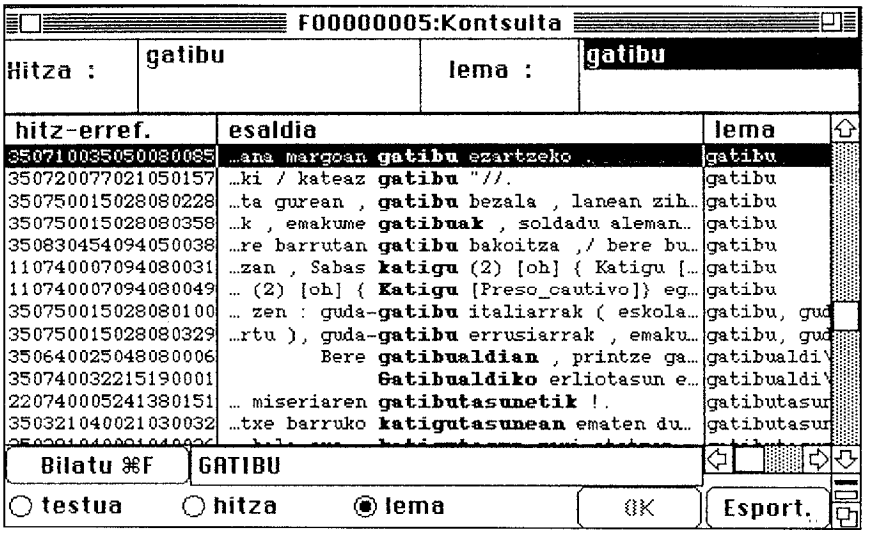
En 1987, UZEI a entrepris un nouveau projet : EEBS (Collecte Systématique du Basque Actuel), corpus de 3 000 000 de mots-textes renouvelé et complété chaque année. Ce projet a été envisagé comme activité complémentaire et plus restreinte au domaine lexicographique. Son but spécifique est de faire connaître l'usage linguistique réel pour ensuite tirer des conclusions sur l'évolution des formes, sur les lacunes à combler, etc.

Pour y parvenir, cinq étapes successives ont été franchies :

1. l'inventaire de ce qui a été publié en basque au XX^{e} siècle (livres, revues, journaux) et la classification de chaque unité d'après quatre paramètres : la période (3 groupes), la distribution géographique (6 dialectes), la typologie du texte (14 types), la masse textuelle (5 groupes) ;
2. un échantillon de 3 000 000 de mots-textes a été obtenu selon des critères statistiques pour qu'il soit aussi représentatif que possible de l'univers inventorié (de quelque 300 millions de formes) ;
3. la mise sur support magnétique des textes statistiquement délimités a été facilitée par le lecteur optique lorsque la qualité du texte le permettait ;
4. le dépouillement des mots-textes a été réalisé automatiquement à l'aide de l'outil *RTerm* ;
5. la lemmatisation, par contre, n'a pu être que semi-automatique, pour des raisons que nous expliquerons plus tard.

Nous avons commencé l'élaboration du corpus sélectionné sur l'outil *RTerm*, développé par HIZKIA (Bayonne) pour le dépouillement systématique de textes et pour la lemmatisation, l'édition des listes et l'obtention de statistiques. En voici la carte de présentation :





Deux remarques s'imposent sur la typologie et la situation sociolinguistique du basque ; elles expliquent la difficulté d'une lemmatisation automatique. Un mot-texte peut se présenter en flexion par le déterminant, le pluralisateur et le cas de déclinaison (par ex. *gatibuak* («le captif») est composé de la forme lexicale *gatibu*, le déterminant *a* et la marque de l'ergatif *k*) ; il peut être aussi l'un des éléments d'un mot composé ou lexie complexe (cf. *guda-gatibu*), etc. Ajoutons à cela une très large dispersion dialectale et sous-dialectale des formes à lemmatiser due, entre autres, à la situation historique particulière du basque, longtemps tenu éloigné de l'enseignement officiel.

Entre 1987 et 1992, UZEI a ainsi créé une base de données de 3 millions de mots-textes lemmatisés. Ce corpus va être renouvelé et complété chaque année, afin de pouvoir refléter fidèlement le basque actuel.

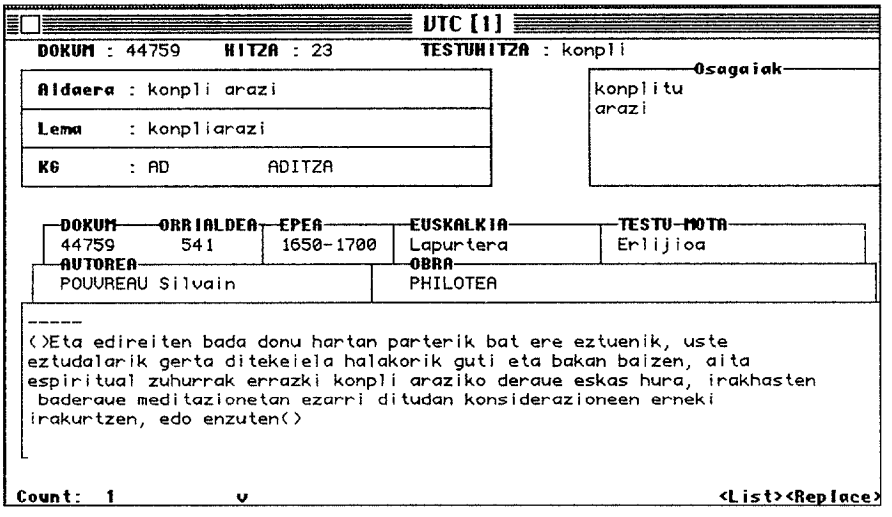
Pour sa part, l'Académie de la langue basque (Euskaltzaindia) a créé, en 1992, la Commission du Dictionnaire Unifié et a demandé à UZEI de travailler pour cette commission, chargée tout d'abord de définir un dictionnaire orthographique, comme premier pas vers la normalisation. Afin de fournir à la commission toutes les données significatives pour son travail, UZEI a obtenu une copie du corpus du *Dictionnaire Général Basque* qui, grâce à ses 5 800 000 entrées, vise à refléter l'usage classique et traditionnel du basque. Ce deuxième corpus a été mis dans un nouveau système standard, ORACLE.

ORACLE est une base de données relationnelle installée sur un système ouvert UNIX. Nous utilisons trois modules principaux : *TextRetrieval* (pour l'exploitation des textes : dépouillement, indexation, recherches) ; *SQLforms* (communication base de données — utilisateur : lemmatisation) et *ProC* (SQL + langage C : programmes et statistiques).

Étant donné que le corpus du *Dictionnaire Général Basque* n'est pas préalablement lemmatisé, nous sommes obligés de lemmatiser les formes sur lesquelles la commission nous demande de l'information (la lemmatisation du corpus dans sa totalité reste donc à faire).

En ce qui concerne la lemmatisation à proprement parler, le nouveau système permet d'établir non seulement le lemme standard, mais aussi le lemme de chaque variante, ainsi que sa catégorie grammaticale et les éléments des mots composés.

C'est pour cela que nous avons jugé que le moment était venu de transférer le corpus EEBS de l'outil *RTerm* à la nouvelle base de données ORACLE qui nous permet le traitement des variantes.



Dès que l'on a lemmatisé les formes à étudier, on a accès directement à l'histoire et à la distribution de chaque forme et de ses variantes. La banque de données est alors prête à fournir des renseignements sur le lemme standard (*lema*), le lemme de la variable (*aldaera*), les éléments des mots composés (*osagaiak*), la catégorie grammaticale (*KG*) de chaque mot-texte, ainsi que sur sa localisation temporelle et géographique grâce aux indications de l'auteur (*autorea*), l'œuvre (*obra*), la période (*epea*) et le dialecte (*euskalkia*), sans compter le contexte.

La base de données permet également d'obtenir des cadres synoptiques présentés selon différents critères. Nous n'en retiendrons ici que les deux plus significatifs.

Synopsis de la distribution d'après les auteurs (simplifié) :

— OEHko ALDAERA DESBERDINEN ERABILPENA —
Lema, aldaera
Epea / Euskalkia / Obra / Autorea
konplimendu 75
konplimendu 62

1	L	TESTAMENTU BERRIA	LEIZARRAGA	Lç	22
1	L	OTHOITZA + KATEXISM.	LEIZARRAGA	Lç *Ins*	9
1	L	ABC	LEIZARRAGA	Lç	3
1	BN	LINGUAE V. P.	Bernard ETXEPARE	E	2
2	L	ELIÇARA	ETXEBERRI Ziburuk.	EZ *Eliç*	1
2	L	GERO	AXULAR	Ax	2
2	L	DEBOZINO ESKUARRA	Jean de HARANBURU	Harb	4
2	L	DOKTRINA	MATERRE	Mat	3
3	L	IMITAZIONEA	Silvain POUVREAU	SP *Imit*	2
3	L	PHILOTEA	Silvain POUVREAU	Sp *Phil*	2
3	Z	ARIMA	TARTAS	Tt *Arima*	1
6	L	MEDITAZIONEAC URTE ...	JAURETXE	Jaur	1
6	BN	BERTSOAK I	BORDEL	Bordel	1
7	L	FABLEAC	GOIETXE	Gy	3
8	L	IPUINAK	BARBIER	Barb *Leg*	1
8	BN	BERTSOAK	ETCHAMENDY	Etcham	1
9	L	MENDEKOSTE GEREZIAK	ETXEPARE JB	JEtchep	2
9	L	IRU ZIREN	LARZABAL	Larz *Iru*	2
konplimentu 1					
6	Z	ASTOLASTERRAK XIX	XX-6	AstLas	1
konprimentu 1					
8	B	ABARRAK II	KIRIKIÑO	Kk *Ab II*	1
kunplimendu			1		
8	BN	BERTSOAK	ETCHAMENDY	Etcham	1
kunplimentu 9					
7	G	GABON GAU BAT	Alfonso ZABALA	Zab *Gabon*	1
7	G	DAMUBA GARAIZ	Justo MOKOROA	Moc *Damu*	1
8	G	EUSKALDUNAK	ORIXE	Or *Eus*	2
8	G	OROITZAK	IRAOLA	Iraola	1
9	G	JOANAK-JOAN	Jon ETXAIDE	Etxde *JJ*	3
9	EB	TOBERAK	Gabriel ARESTI	Arti *Tobera*	1
kunprimentu 1					
8	G	EUSKAL JOKOAK	XX-15	EusJok	1

Synopsis des emplois diachronique et dialectal (simplifié):

	DGV		>>konplitu<< 1-7 (<1900)			8-9 (>1900)				
	B	G	IE	ZuAm	B	G	IE	ZuAm	EB	GUZT
konplimendu	—	2	55	2	1	7	7	—	1	75
konplimendu	—	—	55	1	—	—	6	—	—	62
konplimentu	—	—	—	1	—	—	—	—	—	1
konprimendu	—	—	—	—	1	—	—	—	—	1
konplimendu	—	—	—	—	—	—	1	—	—	1
konplimentu	—	2	—	—	—	6	—	—	1	9
konprimentu	—	—	—	—	—	1	—	—	—	1

Le portrait d'une forme, avec sa tradition, ses déviations, l'emploi diachronique et l'emploi dialectal, sa distribution d'après les auteurs, etc., devient un instrument de travail fondamental pour la Commission du Dictionnaire Unifié qui, en «lisant» ces données-là, pourra en tirer les conclusions pertinentes pour la normalisation et la standardisation de la langue.

Nous avons développé aussi une base de données lexicographique destinée à formaliser et à garder, dans un format abrégé et codé, toute l'information historique et dialectale sur chaque forme (cf. *erabilerak*), ainsi que les critères utilisés par la commission dans la lecture qu'elle fait de cette information (cf. *azalpena*). Cette base de données pourra fournir une base solide pour l'élaboration du dictionnaire normatif de la langue.

Contenu d'un enregistrement:

Format	Kodea	Erabakia	Data
erretore	Z1 : BatHizt	B	94.05.26

Erabilerak

201 :
2021 : 'apaiza' adierakoa bakarrik agertu da: *erretor* da forma nagusia 408 agerraldirekin (IE: 399; B-G: 8; EB: *1*);*erretore* 141 aldiz azaldu da (B: 5; G; 106, IE: 24; B: 6); *ertor* (IE: 35) bakarrik.
2022 : 'eliza bateko buru den apaiza' adieran *erretor* 78 aldiz jaso da (B: 1; G: 5 ; IE: 66; EB: 6), *erretore* 71 (B: 2; G. 21 aldiz (IE): 19, hiru testutan; EB: 20; EgAs: 9) eta *ertor* 3 aldiz (IE). 'Unibertsitateko burua' adierarekin *errektore* eta eratorriak *(errektoregai, errektoretza)* dira nagusi, 13 ager.ekin; eta ager. banarekin: *erretore, erretore-orde, erretoreordetza, erretoretza.*

Azalpena

3121 : *erretore* da hobestekoa (hots: 'parrokiako arduradun nagusia' adieran), B eta G batetik eta IE bestetik baitabiltza; hori da, gainera, erabilera zainduetan hedatuena (ik. 3133 irizpidea ere).
454 : -or / -ore.
5332 : parrokiako arduradun nagusia; cf. *errektore* 'Unibertsitateko kargudun nagusia'.

Iritzi-emaileak

Oharra

jatorrizkoan : *erretor(e).*

Avant de finir cette présentation sommaire des projets d'UZEI dans le domaine de la lexicographie assistée par ordinateur, nous voudrions revenir sur les difficultés de lemmatisation évoquées plus haut. La lemmatisation «manuelle» ou semi-automatique demande un grand effort que nous voudrions alléger et c'est dans ce but que nous voulons créer EUSLEM, un lemmatiseur automatique basé sur l'analyse morphologique du basque (basé, à son tour, sur la morphologie de deux niveaux, cf. Koskenniemi 1983). C'est un projet à développer en collaboration avec la Faculté d'Informatique de l'Université du Pays Basque comme suite à l'analyseur et au correcteur d'orthographe.

Nous présentons ici un schéma des cas auxquels doit faire face le programme :

Forme normalisée	Forme erronée	Forme étudiée par le lexique général (lemme de la variable)	Fautes typiques (règles)	Variable/fautes typiques étudiées par le lexique	Exemple (mot-texte, lemme standard, lemme de la variable, catégorie grammaticale)
+		+			(testuko testu testu IZE)
+		-		-	(afasiagatik ? afasia IZE)
-		-		+	(baltzaren beltz baltz ADJ)
-		-		-	(axtelehenean ? axtelehen IZE)
-		-	+	-	(zuaitzeko zuhaitz zuaitz IZE)
	+	-		-	(extean ? exte IZE)*
+		-		-	(LKB ? LKB 7)

Dans sa configuration actuelle, ce lemmatiseur devrait être capable d'étudier les formes simples. Pour qu'il puisse traiter des formes complexes, une nouvelle démarche s'impose ; elle doit commencer par l'élaboration d'une typologie de ces formes complexes, mais nous n'en sommes qu'aux rudiments.

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

ADURIZ, I., ALEGRIA, I., ARRIOLA, J. M., ARTOLA, X., DIAZ DE ILARRAZ, A., EZEIZA, N., GOJENOLA, K. et M. MARITXALAR (1995) : «Different Issues in the Design of a Lemmatizer / Tagger for basque. "From Text to Tag"», *SIGDAT*, EACL.

ADURIZ, J. A. et M. URKIA (1994) : «Hiztegi Batuaren Datu-Base Lexikografikoa», *Euskera* (Actes du Congrès d'Euskaltzaindia), Leioa, Euskaltzaindia.

AGIRRE, E., ALEGRIA, I., ARREGI, X., ARTOLA, X., DIAZ DE ILARRAZ, A., SARASOLA, K. et M. URKIA (1989) : «Aplicación de la morfología de nos niveles al euskara», *SEPLN*, vol. 8, Barcelone, pp. 87-102.

ALDEZABAL, I., ALEGRIA, I., ARTOLA, X., DIAZ DE ILARRAZ, A., EZEIZA, N., GOJENOLA, K., ADURIZ, I. et M. URKIA (1994) : «EUSLEM: Un lematizador / etiquetador de textos en euskara, *SEPLN*, Córdoba.

DA COSTA, A. (1989) : *RTerm2*, HIZKIA, Bayonne.

Euskaltzaindia (1985, 1990) : *Euskal Gramatika: Lehen urratsak (I, II eta III)*, Bilbo, Euskaltzaindia.

Euskaltzaindia (1992) : *Euskaltzaindiaren Gomendioak eta Erabakiak (I eta II)*, Bilbo, Euskaltzaindia.

KOSKENNIEMI, K. (1983) *Two-Level Morphology: A general Computational Model for Word-Form Recognition and Production*, University of Helsinki, Department of General Linguistics, Publication 11.

Oracle Corporation (1993) : *ORACLE SQL*Forms, SQL*Plus, PL / SQL, Text*Retrieval and so on*, Redwood City.

SAGARNA, A. et M. URKIA (1991) : «Terminología y Lexicografía asistida por ordenador. La experiencia de UZEI», *SEPLN*, vol. 8, Donostia.