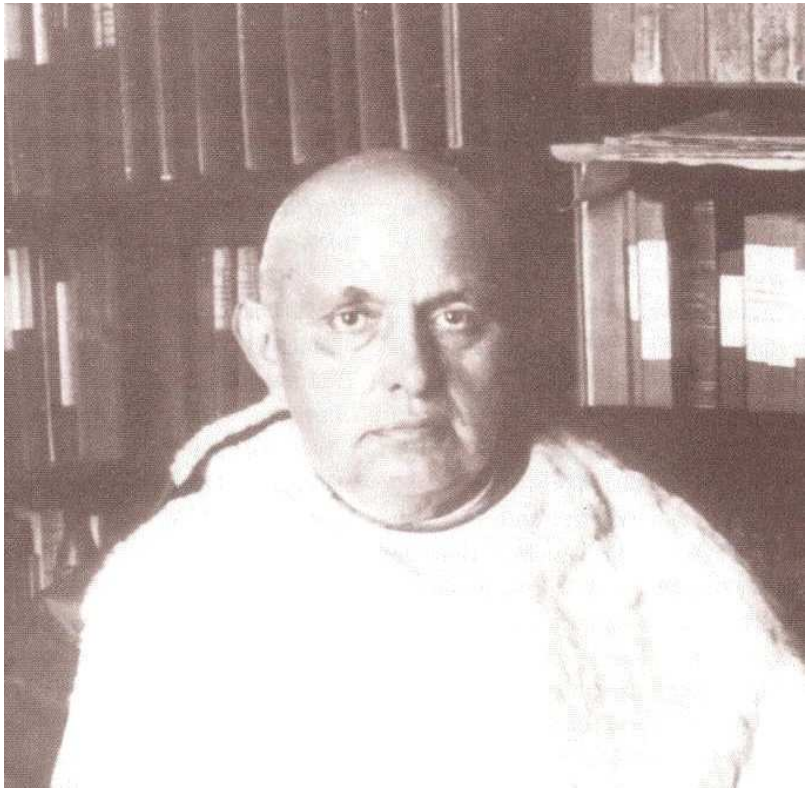
Método general de lematización con una gramática mínima y un diccionario óptimo. Aplicación a un corpus escrito dialectal

Hiroto Ueda (Universidad de Tokio)
uedahiroto@jcom.home.ne.jp
Maria-Pilar Perea (Universitat de Barcelona)
mpilar.perea@ub.edu

1. Descripción del corpus

- Antoni M. Alcover (Manacor 1862 - Palma 1932)
- *Diccionari català-valencià-balear*
- *Bolletí del Diccionari de la Llengua Catalana* (BDLC)
- *La Aurora*



1. Descripción del corpus

- BDLC


Butlletí del Diccionari de la Llengua Catalana

Suscripció: 2 pess. en l' any.—Redacció: Serra, 13.—Palma de Mallorca

- 14 volúmenes (1901-1926)

1. Descripción del corpus

Informaciones contenidas en el *BDLC* (primera época):

1) textos, artículos y estudios del propio Alcover (temas filológicos, dialectales, onomásticos, históricos, lingüísticos, reseñas bibliográficas, necrologías, los dietarios y los manifiestos);
2) textos y artículos de carácter filolológico Francesc de B. Moll, que colabora en el *Bolletí* desde el volumen XIII;
3) artículos más o menos extensos de escritores, lingüistas y eruditos de la época;
4) artículos periodísticos publicados en la prensa, que ilustran acontecimientos culturales;
5) las observaciones dialectales —«Notes dialectals»— que enviaban algunos corresponsales y colaboradores de la *Obra del Diccionari*.

Per consegüent ... dins l' Introducció ...
matica y dins la *Fonètica* dexa sense cap demostració ...
primera tesis: *La llengua primitiva de Catalunya* ...

Col·laboradors qui tenen poca son

En Pere Basté, de Barcelona, ha omplides ... dal-s.—En Victori Santamaría, notari de Vendrell ... En Joan Climent, d' Agullana (Alt-Ampurdà), ... Bernadí Mateu, de Mancor (Mallorca), 1000 de bell ... Francesch Monserrat, Vicari de Ca 's Concos (Mallorca) ... —En Josep Arqués y Arrufat, seminarista de Barcelona ... —Moss. Josep Valls, professor de Llatí, de Prats de ... 2200;—Moss. Jaume Pons, de Bunyola, 1000;—Moss. ... de Sampedor (Prov. de Barcelona), 1076;—Moss. ... Bosch, advocat, de Les Gunyoles (Panadés), ... Vallés y Vidal, de Barcelona, 1070;—En Ramon ... Igualada, 1000;—Moss. Jaume Pascual, Vicari d' A... 7000 de bell nou;—En Fidel de Moragas, de Valls, 6...

Crònica de l' Obra del diccionari

Els dos articles sobre la llengua ibèrica se son allargats ... que no 'ns dexen espay per les altres seccions. Axí ... porem donar conte fins a n-el *Bolletí* de desembre d' ... de coses qu'interessen a n-els col·laboradors, referents a ... *del Diccionari*, que, gràcies a Deu, va en popa de tal ... mos ho hauríem pensat. Cada dia s' hi afigen col·laboradors ... vells, y aumenta la suscripció del *Bolletí*, y sovintetgen ... dels amichs qu'ens fan a sebre el milenars de cèdules ... plides, y la *taula d' honor* dels *qui tenen poca son*, ... qu' es un goig de Deu.

No res, endevant les atxes!

PALMA DE MALLORCA—Tipo-litografia de ...

Bolletí del Diccionari de la Llengua Catalana

Suscripció: 2 pess. en l' any.—Redacció: Serra, 18.—Palma de Mallorca

SUMARI: La nostra anada a Barcelona.—Un congrés gramatical.—Crònica de l' Obra del Diccionari.—Col·laboradors qui tenen poca son.—[illegible]

La nostra anada a Barcelona

Hi arribàrem dia 27 de novembre.

Sempre 'ns hi campam bé, gràcies a Deu, a n-aquexa gran ciutat, y no param de rebrer mostres extraordinàries de consideració, afecte coral y viu entusiasme per *l'Obra del Diccionari*, que may sebrem agrair axí com cal.

Donarem conte de lo que hi hem fet aquesta vegada referent a la nostra *Obra* seguint l' orde dels dies que hi fórem.

Dia 27

A les cinch del cap vespre donàrem una conferencia a n-el *Centre Moral y Instructiu* de Gràcia sobre 'l tema: *Importancia moral y instructiva del Diccionari Català*.

Vetassí que 'n diu el *Diario de Barcelona* de dia 29 de dit mes: «—El M. Iltre. Sr. D. Antonio María Alcover, Vicario general de Mallorca, dió en la noche del domingo último una conferencia en el Centro Moral Instructivo de Gracia, en la que trató de la importancia del Diccionario Catalán como obra moral é instructiva. El conferenciante, llegado el mismo día de Palma, expresó ante todo la emoción que le produjo la hermosa manifestación de fe que había presenciado aquella mañana en las calles de Barcelona, elogió luego los trabajos del Centro, que dijo conocer de muchos años y, entrando en el asunto de su discurso, ponderó la transcendencia de la obra del Diccionario, desde el punto de vista moral, refiriéndose para probarlo á la encíclo-

- Digitalización del BDLC

- Publicación en CD-ROM en dos ediciones (2003) y (2004)

BDLC

Tomo/núm.	Año	pág.	palabras	Tomo/núm.	Año	pág.	palabras
I (17)	1902-1903	587	228.518	VIII (8)	1914-1915	268 +114	157.298
II (13)	1904-1905	408	138.358	IX (13)	1916-1917	384	145.240
III (9)	1906-1907	414	165.719	X (13)	1918-1919	524	217.883
IV (11)	1908-1909	405	176.599	XI (4)	1920	336	125.158
V (dietario)	1908	378	167.600	XII (6)	1921-1922	368	153.293
VI (21)	1910-1911	392	171.785	XIII (7)	1923-1924	376	168.496
VII (13)	1910-1911	436	183.627	XIV (5)	1925-1926	352	121.805

2. PROCESOS DE LEMATIZACIÓN

2.1. Texto

22 (...) Bo es de veure que quedam amichs. Mostra a l'alemanya unes castanyetes noves ab uns grans flochs y borlins. —Això no es de Catalunya, li dich. —No, diu ell, es d'Andalusia. Si un estranger se'n vol dur res característich d'Espanya, compra unes castanyetes; si de Catalunya, una barretina y unes espardenyes. —Prou que hi ha molts d'espanyols que donen peu a formar tals judicis d'Espanya.

Texto. 1. Texto llano (un fragmento)

2.2. Lista de Lema-Formas

LEMA	Cat.	Otras formas
a	P	
això	M	
alemany	S+A	
amb	P	
amich	S	
anar	V	vaig, vas, va, an, vag
Andalusia	E	
aprendre	V	aprend
aqueix	M	aquexos, aqueixa, aqueixes

Fig. 1. Lista de Lema-Formas

&	LEMA
a	_P_a
ab	@
això	_X_això
alemanya	_A_alemany
amichs	_S_amich
andalusia	_E_Andalusia
barretina	_S_barretina
bo	_A_bo
borlins	_A_bo

Fig. 2. Lista del resultado

22 (...) Bo_A es_T_V_X de_P veure_V que_C quedam_V amichs_S. Mostra_S_V a_P l_T_X’alemanya_A unes_T castanyetes_S noves_A ab @uns_T grans_A flochs_S y_C borlins_A. —Això_X no_D es_T_V_X de_P Catalunya_E, li_X dich_V. —No_D, diu_V ell_X, es_T_V_X d_P’Andalusia_E. [...]

Texto. 2. Texto asignado de Categorías

2.3. *Casos ambiguos y desambiguación*

1) Delante de un sustantivo (S), _A_X (adjetivo / pronombre) debe convertirse en _A (adjetivo), por ejemplo: tals_A_X judicis_S

2) Delante de un sustantivo (S), _T_X (artículo / pronombre) debe convertirse en _T (artículo), por ejemplo: sebre’l_T_X castellà_S

3) Delante de un verbo (V), TX (artículo / pronombre) debe convertirse en _X (pronombre), por ejemplo: per que el_X vejen_V

1) (&)_A_X(@&_S)=>1_A2
2) (&)_T_X(@&_S)=>1_T2
3) (&)_T_X(@&_V)=>1_X2

22 (...) Bo_A es_V de_P veure_V que_C quedam_V amichs_S. Mostra_V a_P l_T' alemanya_A unes_T castanyetes_S noves_A ab_P uns_T grans_A flochs_N y_C borlins_S. —Això_M no_D es_V de_P Catalunya_E, li_X dich_V. —No_D, diu_V ell_X, es_V d_P' Andalusia_E. [...]

Texto. 3. Texto desambiguado

2.4. Lematización

22 (...) Bo_A_bo es_V_ésser de_P_de veure_V_veure que_C_que quedam_V_quedar amichs_S_amich. Mostra_V_mostrar a_P_a l_T_el' alemanya_A_alemany unes_T_un castanyetes_S_castanya noves_A_nova ab_P uns_T_un grans_A_gran flochs_N y_C_i borlins_S. —Això_M_això no_D_no es_V_ésser de_P_de Catalunya_E_Catalunya, li_X_li dich_V_dir. —No_D_no, diu_V_dir ell_X_ell, es_V_ésser d_P_de' Andalusia_E_Andalusia. [...]

Texto. 4. Texto lematizado

3. EXPERIMENTO Y RESULTADO

	Voces	Formas	Lemas
Cantidad total	19.703	2.625	822
Valor máximo	2.277	2.277	3.512

Tabla. 2. Voces y lemas

V,C,L:Voz	B-05	Acum.
ha	2277	2277
es	1908	4185
son	698	4883
he	636	5519
fa	591	6110
té	383	6493
som	337	6830
veure	307	7137
està	246	7383
fan	217	7600

Fig. 3. Voces

V,C,L:LEMA	B-05 (2)	Acum
haver	3512	3512
ésser	3483	6995
fer	1260	8255
tenir	730	8985
anar	657	9642
veure	625	10267
estar	475	10742
dir	410	11152
trobar	398	11550
poder	289	11839

Fig. 4. Lemas

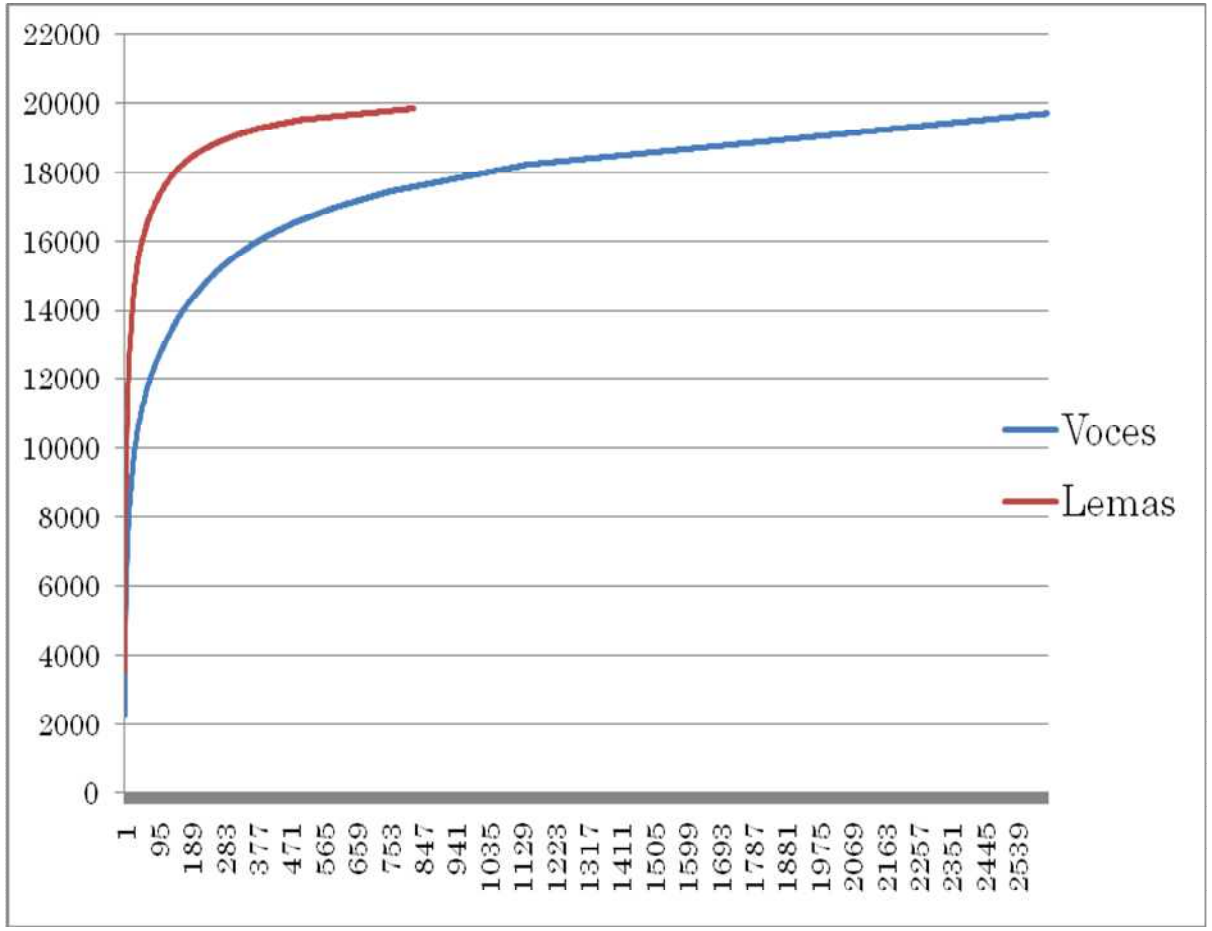


Fig. 5. Voces y Lemas

4. CONCLUSIÓN

En la lingüística se persiguen la precisión, la concisión y la exhaustividad. En la descripción de una lengua, para llegar al nivel deseado, se tiene en cuenta el estado tanto del Diccionario como de la Gramática y no se consideraría satisfactoria una precisión de un 95%. La cuestión de tratamiento lingüístico del texto no obstante no es preparar un diccionario gigantesco ni una gramática escrupulosa sin necesidad. Nuestra propuesta es buscar un punto equilibrado donde colaboren un diccionario óptimo y una gramática mínima.

REFERENCIAS

Chrupała, G. (2006). “Simple data-driven context-sensitive lemmatization”. *Proceedings of SEPLN.* http://www.sepln.org/revistaSEPLN/revista/37/16.pdf (2010/ 2/8)

Gómez Díaz, R. (2005). *La lematización en español: una aplicación para la recuperación de información.* Gijón: Ediciones Trea.

Guirao, J. M. & Moreno-Sandoval, A. (2004) “A “toolbox” for tagging the Spanish C-ORAL-ROM corpus”. *IV*

International Conference on Language Resources and Evaluation (LREC2004) Proceedings. http://lablita.dit.unifi.it/coralrom/papers/ toolbox-final.pdf (2010/2/8).

Loftsson, H. (2008). “Tagging Icelandic text: A linguistic rule-based approach”. *Nordic Journal of Linguistics*, 31.p.1-29. http://www.ru.is/faculty/hrafn/Papers/ IceTagger_final.pdf (2010/2/8).

McEnery, T. (2003). “Corpus linguistics”. En R. Mitkov (ed.), *The Oxford Handbook of Computational linguistics*. Oxford: Oxford University Press, 448-463.

Mueller, M. (2009). “NUPOS: A part of speech tag set for written English from Chaucer to the present”, http://panini.northwestern.edu/mmueller/nupos.pdf.

Meyer, C. F. (2002). *English corpus linguistics. An introduction.* Cambridge: Cambridge University Press.

Megyesi, B. (2002). “Shallow Parsing with PoS Taggers and Linguistic Features”, *Journal of Machine Learning Research*, 2, 639-668. http://jmlr.csail.mit.edu/papers/volume2/megyesi02a/megyesi02a.pdf (2010/2/10).

O’Donovan, B. & Troussov, A. (2003). “Morphosyntactic annotation and lemmatization based on the finite-state dictionary of wordformation elements”. *Proceeding of the*

International Conference Speech and Computer, Moscow, Russia.
http://www.iol.ie/~bodonovan/pubs/SPECOM-03.pdf.

Perea, M.-P. (ed.) (2003). *Bolletí del Diccionari de la Llengua Catalana*. Palma: Conselleria d'Educació i Cultura. Govern de les Illes Balears. CD-ROM edition.

Perea, M.-P. (ed.) (2004). *Bolletí del Diccionari de la Llengua Catalana* nova edició ampliada amb índexs). Palma: Conselleria d'Educació i Cultura. Govern de les Illes Balears. CD-ROM edition.

Plisson, J., Lavrac, N. & Mladenic, D. (2004). "A Rule based Approach to Word Lemmatization". *Proceedings of the 7th*

International Multi-Conference Information Society, 1(1):83-86. http://eprints.pascal-network.org/archive/00000715/01/Pills on- Lematization.pdf (2010/2/8).

Samuelsson, C. & Voutilainen, A. (1997). “Comparing a Linguistic and a Stochastic Tagger”. *8th Conference of the European Chapter of the Association for Computational Linguistics*, *Proceedings of the Conference*, Madrid: UNED, 246-253. http://www.aclweb.org/anthology/P/P97/P97-1032.pdf (2010/2/10).

Siemens, R. G. (1996). “Lemmatization and parsing with TACT

preprocessing programs". *Computing in the Humanities Working Papers*. http://www.chass.toronto.edu/epc/chwp/siemens2/index.html (2010/2/10).

Ueda, H. (2005). "Methods of 'hand-made' corpus linguistics - A bilingual data base and the programming of analyzers". *Usage-Based Linguistic Informatics 1, Linguistic Informatics -State of the Art and the Future*. John Benjamins Publishing Company, 145-166.

Voutilainen, A. (2003). "Part-of-speech tagging". In R. Mitkov (ed.) *The Oxford Handbook of Computational linguistics*. Oxford University Press, 219-232.