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

HIROTO UEDA
*Universidad de Tokio*
MARIA-PILAR PEREA
*Universitat de Barcelona*

*Resumen*
*A falta de lematizadores y etiquetadores para lenguas minoritarias proponemos elaborar nuestro propio procesador programado por Microsoft Excel VBA. El proceso consiste en asignar provisionalmente la categoría de partes de oración por un diccionario preparado en procesos anteriores y desambiguar los homógrafos por una lista de reglas gramaticales, también almacenadas anteriormente, para proceder finalmente a la lematización utilizando el mismo diccionario ahora dotado de información ortográfica y gramatical (categoría).*
*Este método es flexible y aplicable a distintas lenguas europeas y consigue ofrecer un resultado cada vez mejor a medida que en cada operación se nutren tanto el diccionario como la gramática. Nuestra idea es crear un aparato de procesamiento común, que se activa con parámetros léxicos y gramaticales específicos de cada lengua, objeto de investigación.*
*Palabras clave: lematización, Microsoft Excel VBA, catalán, lengua minoritaria, diccionario, gramática*
*Abstract*
*Owing to the lack of lemmatizers and taggers for minority languages, we propose to develop our own processor programmed by Microsoft Excel VBA. The process consists in assigning temporarily the category of parts of sentences through a dictionary prepared in previous processes and disambiguating homographs using a list of grammatical rules, also stored previously, in order to finally lemmatize the text using the same dictionary now provided with spelling and grammar (category) information.*
*This method is flexible and applicable to different European languages and offers a better result as both the dictionary and the grammar are fed on each transaction. Our idea is to create a common processing apparatus, which is activated with language specific lexical and grammatical parameters.*
*Keywords: lemmatization, Microsoft Excel VBA, Catalan, minority languages, dictionary, grammar*

## 1. INTRODUCCIÓN

Uno de los procesos fundamentales para tratar los textos lingüísticos es la lematización, que consiste en asignar una forma representativa a distintas formas concretas variables: formas conjugadas del verbo, cambios según el género y número de adjetivos y sustantivos, etc. Es necesario, a la hora de realizar unos estudios estadísticos, redactar un vocabulario o diccionario, intentar una búsqueda de información por palabras clave y analizar la combinación de elementos, entre otros objetivos de investigación.[1]

[1] Para la información general de anotaciones del texto lingüístico, véanse por ejemplo Meyer (2002) y McErney (2003). Contamos con una monografía sobre el tema de lematización del español en la obra de Gómez Díaz (2005).

Aparte de los procesadores existentes para lenguas mayoritarias, como el inglés, el español, el japonés, etc., los investigadores de lenguas minoritarias se ven obligados a efectuar la lematización manualmente. Para salvar esta dificultad, vamos a proponer un método sencillo, que podría ser aplicable a distintas lenguas europeas, junto con algunas consideraciones teóricas y prácticas. Nuestro objeto de lematización es el corpus preparado en el proyecto de edición en CD-ROM del *Bolletí del Diccionari de la Llengua Catalana* (2002-2004).[2]

## 2. Descripción del corpus

En dialectología catalana es bien conocida la figura de Antoni M. Alcover (Manacor 1881 - Palma 1931). Gestor y redactor del *Diccionari català-valencià- balear*, emprendió prolongadas encuestas dialectales, recopiló narraciones populares y, entre otras muchas actividades, redactó casi íntegramente dos publicaciones que consiguieron una muy buena acogida en su época: la revista titulada *Bolletí del Diccionari de la Llengua Catalana* (BDLC) y el semanario *La Aurora*, editado en su villa natal, Manacor.

El *Bolletí del Diccionari de la Llengua Catalana* se considera la primera revista de carácter filológico publicada en Cataluña y en el Estado español. Alcover la creó para que se convirtiera en la tribuna de divulgación y de propaganda de su *Obra del Diccionari i de la Gramàtica*, nombre con que se conocía el proyecto de elaboración del diccionario citado anteriormente.

Bajo la dirección de Alcover, se publicaron catorce volúmenes del BDLC (entre el 1901 y el 1926), y, a pesar de que contó con colaboraciones más o menos esporádicas de distintos eruditos, la mayor parte de la redacción fue obra exclusiva del dialectólogo. La revista fue concebida inicialmente como una publicación mensual de 16 páginas, pero a menudo la puntualidad se incumplía, y el lector recibía números mucho más extensos aunque aparecieran dos o tres meses después. En general, la unión de los números de dos años ha dado lugar a un tomo (Tabla 1), excepto en el tomo V, que corresponde a la publicación de la primera excursión de Alcover a países europeos, y el XI, que agrupa exclusivamente los números publicados en 1920.

A grandes rasgos, el *BDLC*, en su primera época, contiene las informaciones siguientes:

---

[2] El proyecto fue financiado por la Conselleria d'Educació i Cultura del Govern de les Illes Balears. Véase Perea (2003, 2004).

1) de manera mayoritaria, textos, artículos y estudios del propio Alcover, que tratan, por un lado, de temas filológicos, dialectales, onomásticos, históricos, lingüísticos, y reúnen, por otro lado, numerosas reseñas bibliográficas, necrologías, los dietarios y los manifiestos;

2) textos y artículos de Francesc de B. Moll, que colabora en el *Bolletí* desde el volumen XIII, y se convierte en su editor en 1933, iniciando la segunda y última etapa de la revista, con la publicación del volumen XV;

3) artículos más o menos extensos de escritores, lingüistas y eruditos de la época;

4) artículos periodísticos publicados en la prensa de la época, que ilustran acontecimientos concretos, como la realización del Primer Congrés Internacional de la llengua catalana, en 1906, o las diversas manifestaciones del movimiento de recuperación de la lengua catalana que se llevaron a cabo a inicios del siglo XX;

5) las observaciones dialectales —las llamadas «Notes dialectals»— que enviaban algunos corresponsales y colaboradores de la *Obra del Diccionari*.

El contenido del BDLC se ha publicado en CD-ROM en dos ediciones (2003) y (2004). Y ha sido de ésta última de donde se ha seleccionado el corpus que se pretende estudiar.

Para nuestro estudio, que se centra en el corpus escrito alcoveriano, únicamente se toma en consideración el contenido del apartado (1). Las otras informaciones se omiten, puesto que no corresponden al mismo autor.

Se indica a continuación la extensión de cada tomo, los números que contiene, el año de publicación y el número de palabras, una vez que se han eliminado las informaciones no alcoverianas.

Tabla 1: Descripción formal del *Bolletí del Diccionari de la Llengua Catalana* (BDLC)

| Tomo/núm. | Año | pág. | palabras | Vol. | Año | pág. | palabras |
|---|---|---|---|---|---|---|---|
| BDLC I (17) | 1902-1903 | 587 | 228.518 | BDLC VIII (8) | 1914-1915 | 268+114 | 157.298 |
| BDLC II (13) | 1904-1905 | 408 | 138.358 | BDLC IX (13) | 1916-1917 | 384 | 145.240 |
| BDLC III (9) | 1906-1907 | 414 | 165.719 | BDLC X (13) | 1918-1919 | 524 | 217.883 |
| BDLC IV (11) | 1908-1909 | 405 | 176.599 | BDLC XI (4) | 1920 | 336 | 125.158 |
| BDLC V (dietario) | 1908 | 378 | 167.600 | BDLC XII (6) | 1921-1922 | 368 | 153.293 |
| BDLC VI (21) | 1910-1911 | 392 | 171.785 | BDLC XIII (7) | 1923-1924 | 376 | 168.496 |
| BDLC VII (13) | 1912-1913 | 436 | 183.627 | BDLC XIV (5) | 1925-1926 | 352 | 121.805 |

## 3. Procesos de lematización

### *3.1. Texto*

Los textos no presentan siempre un estado ideal para el procesamiento automático, el cual cuenta con reglas gramaticales y ortografía establecida. Ante la realidad complicada de la lengua, lo normal sería tratar tales textos "problemáticos" de forma manual para salvar las dificultades técnicas que puedan presentar algunos tratamientos automáticos. La manipulación manual es necesaria en el caso de algunos datos especiales y no lo es para procesamientos regulares. Cabe distinguir entre el procesamiento de tipo específico y el de tipo general. En esta sección intentaremos presentar nuestro método de lematización de un corpus dialectal a través de algunos experimentos reales y de los resultados obtenidos a través de procesamientos manuales y automáticos.

El Texto 1 muestra un fragmento del texto a que nos hemos referido en la introducción. Cada párrafo está situado en la celda de Excel junto con el número de identificación correspondiente en la celda izquierda. Nuestro propósito es lematizar todas las formas verbales asociándolas con la forma representativa.

| |
|---|
| 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)

A partir del texto llano digitalizado podemos elaborar automáticamente una lista de frecuencia de cada voz con relación al total del texto: *de* (10.191), *y* (8.849), *a* (5.251), *la* (4.332), *l* (3.410), *el* (3.241), *d* (3.139), *un* (2.867), *hi* (2.839), *que* (2.800), *les* (2.415), *una* (2.294), *ha* (2.277), ... hasta multitud de voces de una sola ocurrencia.

Como primer intento, pensamos asignar un lema a cada voz, y aplicar la lista de correspondencia Forma - Lema al texto. Esta lista de correspondencia sería útil para evitar el trabajo repetitivo de asignación de las mismas formas de alta frecuencia. Una lista así preparada serviría también para trabajos posteriores con otros textos del mismo corpus, añadiendo las nuevas voces que aparecen en el nuevo texto. Este método sería la única solución, al carecer de un programa de etiquetado específico, que se han elaborado en mayor

envergadura para las grandes lenguas europeas.[3] Se trata de un método directo con un máximo de diccionario (lista de correspondencia) y sin gramática (categorización y reglas).

En esta ocasión, en cambio, proponemos un método ecléctico buscando un punto óptimo de colaboración entre la gramática y el diccionario. Por no poseer una gramática aplicable al procesamiento de textos, de momento contaremos con un mínimo de información gramatical: la categoría de partes de oración.

*3.2. Lista de Lema-Formas*

Una de las características comunes de las lenguas indoeuropeas es su morfología verbal basada en la conjugación, flexión de las formas terminales. Es decir, cada forma verbal está constituida básicamente por una parte anterior invariable (Raíz) y otra posterior variable (Terminación). Es una regla morfológica sencilla, aparte de los casos de cambios de radicales, por ejemplo, *hago, haces, hice, hecho, etc.* del verbo español *hacer*; o *faig, fas, fa, fem feu, fan* del catalán *fer*; y de supletismo, *voy, vas, va* ... del verbo español *ir* y *vaig, vas, va, anem, aneu, van* del verbo catalán *anar*. Nuestra idea es preparar provisionalmente una bolsa de infinitivos más frecuentes y, a partir de esta bolsa, asignar automáticamente un lema a las nuevas formas que aparecen en el texto eligiendo la forma más parecida posible. Por ejemplo, la voz *abeuren* debe corresponder al lema *abeurar* 'abrevar' por tener una parte común *abeur*.

Con tal de que esté preparada una buena lista de correspondencia Raíz - Lema, en general la mayoría de las veces una forma se identifica con su lema correcto. Ahora bien, nuestro trabajo manual ya no es asignar uno por uno el lema correspondiente a la nueva forma que aparece en el nuevo texto, sino simplemente registrar la nueva forma de la raíz en la lista de raíces:

[3] Veánse los sitios de Brill's Tagger: http://www.tech.plym.ac.uk/soc/staff/guidbugm/pysoftware.htm; CLAWS: http://ucrel.lancs.ac.uk/claws/
TreeTagger: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/; de los cuales Brill's Tagger y CLAWS son etiquetadores específicos de inglés, mientras que Tree Tagger permite uso de varias lenguas europeas.

| 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 |

Figura 1: Lista de Lema-Formas

La Figura 2 muestra una parte del resultado del análisis del Texto:

| & | 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 |

Figura 2: Lista del resultado

donde observamos que *amichs* se ha identificado correctamente con el lema *amich*,[4] aunque la forma plural no figure en la lista. Las palabras que no han sido analizadas por falta de los datos del Diccionario aparecen con el signo de arroba (@), lo cual quiere decir que se debería asignar la categoría y lema en el Diccionario. El caso de "borlins", asignado equivocadamente con "A_bo", se debe a que el programa ha buscado el adjetivo "bo" como último recurso. Con la inclusión de *borlin* en el Diccionario se resuelve este problema.[5]

Nuestro programa también puede ofrecer un texto de resultado. En este caso nos limitaremos a presentar la asignación automática de la categoría gramatical.[6]

[4] La grafia <amich> corresponde a <amic> del catalán estándar.

[5] El caso anterior de *ab* @ es distinto. No se identifica con a (P_a), puesto que la preposición exige una identificación total, mientras que las palabras variables (verbos, adjetivos, sustantivos) permiten una identificación parcial. La identificación de *borlins* con *bo* podría evitarse asignando algunas condiciones especiales de la terminación de adjetivos. De momento, sin embargo, seguimos trabajando sin especificaciones de condiciones específicas para conservar la característica generalizadora del programa, aplicable a distintas lenguas.

[6] Los signos utilizados son: A: Adjetivo, C: Conjunción, D: Adverbio, I: Interjección, M: Demostrativo, N: Numeral, O: Posesivo, P: Preposición, R: Relativo, S: Sustantivo, T: Artículo, V: Verbo, X: Pronombre. Naturalmente estos signos se pueden cambiar según la conveniencia del usuario.

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

*3.3. Casos ambiguos y desambiguación*

Al observar el Texto 2, nos damos cuenta de que algunas Formas poseen más de una asignación, por ejemplo, "es_T_V_X", "l_T_X", "alemanya_S_A", etc. Se trata de casos ambiguos respecto a la categoría, puesto que se ha hecho a partir de la lista de formas, sin tener en cuenta su sintaxis. Precisamente para poder ofrecer estos casos ambiguos, hemos introducido el procesamiento de categorización. El mérito de asignación de categoría gramatical es su capacidad de distinguir entre varios homógrafos: *trabajo* como sustantivo y *trabajo* como primera persona singular del verbo *trabajar* en español; y *sebre* 'saber' (sustantivo) y *sebre* 'saber' (verbo) en catalán dialectal. La mayoría de las veces se distinguen por la categoría, por ejemplo S(ustantivo) y V(erbo): trabajo_S, trabajo_V. A partir de esta asignación por "V", podemos proceder a la lematización verbal, excluyendo los casos de "S" (sustantivo).

Para distinguir entre T (artículo) y X (pronombre), contamos con la información sintáctica siguiente:[7]

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

Para estas Reglas gramaticales, se elaboran las siguientes fórmulas, que se basan en las Expresiones Regulares:[8]

[7] Sin recurrir a las reglas gramaticales, se podría solucionar el problema de ambigüedad por medidas estadísticas, que consisten en buscar la mayor probabilidad posible de secuencias de tres elementos (trigramas) extraídos de textos anotados. Véase Voutilainen (2003).

[8] Estas fórmulas están basadas principalmente en la versión de Expresiones Regulares de Microsoft VBScript, con algunas modificaciones simplificadoras.

1) (&)_A_X(@&_S)=>$1_A$2

2) (&)_T_X(@&_S)=>$1_T$2

3) (&)_T_X(@&_V)=>$1_X$2

donde "&" representa una secuencia de letras utilizadas en las palabras, "@" es una secuencia de letras no utilizadas en las palabras, $1 corresponde a la secuencia de letras entre la primera paréntesis (&) y $2, a la de la segunda paréntesis (@&_S). El signo de "=>" significa que la fórmula izquierda se convierte en la fórmula derecha.

Estas asignaciones se almacenan en la Lista de Reglas, que se utiliza cada vez que se obtiene un texto ambiguo. El resultado es:

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

*3.4. Lematización*

Una vez desambiguado el texto por varias fórmulas gramaticales, se procede finalmente a su lematización. La lematización consiste en asignar el lema correspondiente único. Con la información de la forma de la voz y su categoría gramatical, se identifica con su lema correspondiente, con alta precisión.

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

El proceso general de lematización de un texto consiste en: (1) Primera categorización gramatical automática; (2) Corrección manual de errores del Diccionario; (3) Segunda categorización gramatical automática; (4) Corrección manual de errores de la Gramática; (5) Lematización automática.

En el método propuesto se combina el procesamiento automático y la asignación manual. En cuanto al procesamiento automático, hemos elaborado un programa general que podría ser aplicable a distintos idiomas flexivos de terminación, de tipo indoeuropeo. La diferencia consiste en la formación del Diccionario y en sus propias reglas de la Gramática. La tarea del investigador se divide en dos partes: preparación de un Diccionario óptimo y preparación de una Gramática mínima. Cuánto menor es la cantidad de elementos en el Diccionario, la búsqueda resulta más rápida. En la recopilación de la Gramática, se tiene que tener en cuenta no solamente la formulación adecuada, sino también su orden de aplicación. Aquí también se desea la cantidad mínima de las reglas, para lo cual se debería estar versado tanto en la sintaxis de la lengua como en el manejo de Expresiones Regulares. Lo ideal sería que colaboraran un lingüista y un especialista de ciencias informáticas.

Existen varios proyectos de lematizadores y etiquetadores en el mundo académico de la lingüística de corpus. Algunos son aplicables a solo una lengua y otros, aplicables a varios idiomas, alcanzan apenas 90 o 95% de precisión. Los programas se realizan de manera independiente de nuestro ámbito de trabajo. El método que hemos propuesto difiere de los anteriores en los puntos siguientes: (1) es aplicable a múltiples lenguas europeas; (2) se asciende el grado de precisión a medida que se nutren tanto el Diccionario como la Gramática;[9] (3) los procesos están programados para Microsoft Excel, nuestro ámbito de trabajo de siempre, de modo que hay una buena continuación del lugar de trabajo y de los textos tratados de cantidad relativamente reducida.[10]

## 4. Experimento y resultado

En cuanto a la labor de identificación, en teoría se supone que la lematización por categorías resulta más económica que la de por formas. El proceso mismo de la lematización serviría como una prueba de esta hipótesis. Ahora nos preguntamos qué grado de diferencia existe entre los dos métodos. Para realizar el experimento de comprobación utilizamos dos textos del volumen 5 de BDLC: uno de texto llano donde aparecen voces diferentes, y otro de texto lematizado donde aparecen lemas unificados de las voces. Si nos fijamos en las formas

[9] El estado del resultado es flexible dependiendo de la cantidad de información disponible y de la buena constitución de las reglas. Consideramos que la flexibilidad de uso que permiten los programas es importante a la hora de aplicar a distintas lenguas y a distintos objetivos de investigación. Nuestro método no exige unas etiquetas previamente establecidas ni reglas gramaticales incorporadas, sino que es adaptable a las condiciones del usuario. Para las directrices diferentes del programa, véase Mueller (2009).

[10] Cf. Tabla 1. Para la utilización de programas codificados en macro de Excel, véase Ueda (2005), donde hemos explicado las funciones de nuestro sistema SIAL (Sistema Integral para Análisis Lingüísticos).

verbales, en el corpus aparecen 19.703 voces en total que se distribuyen de la manera siguiente:

Tabla 2: Voces y lemas

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

En la Tabla 2 observamos que hay mucha más cantidad en voces unificadas que en lemas. Si se trabaja manualmente con formas (2.277), el coste de la labor en el tratamiento es tres veces mayor grande que el de lemas (822). Es impensable trabajar con las voces concretas que aparecen en el Texto (19.703).

La cantidad máxima en las voces corresponde a la forma *ha*, que posee el valor de 2.277, mientras que en la de los lemas es del verbo *haver*, 3.512. Las formas y los lemas subsiguientes son los que se presentan en las Figuras siguientes:

| 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 |

Figura 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 |

Figura 4: Lemas

Para observar la curva de las cantidades acumuladas de los elementos en orden decreciente, hemos elaborado la Figura 5:

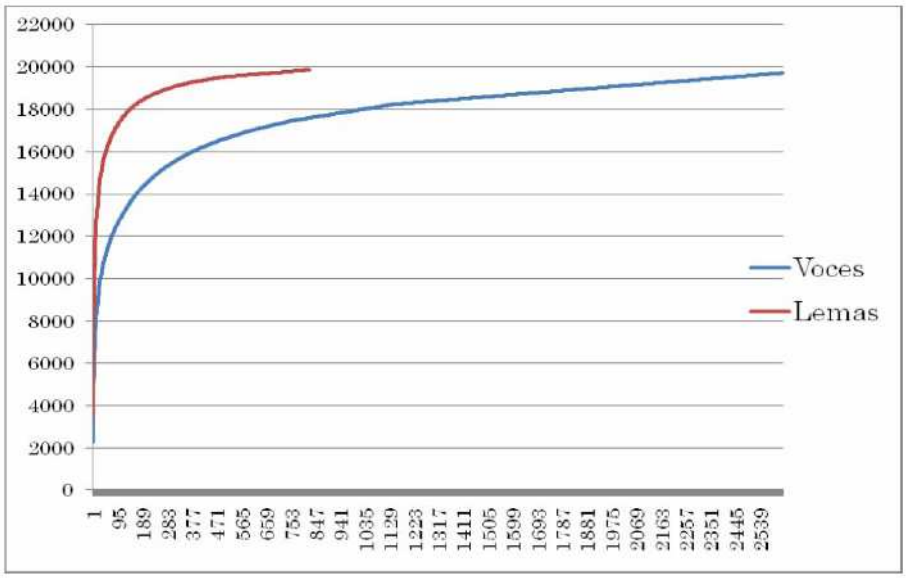


Figura 5: Voces y Lemas

En esta Figura se observa que la curva ascendiente de lemas es considerablemente más destacable que la de las voces. En la práctica de lematización, se supone que en los primeros textos tratados aparecen las voces frecuentes. Una vez registrados las voces y sus lemas en el Diccionario, el coste de los trabajos siguientes se reduce notablemente. Lo mismo puede decirse de la alimentación de la lista de Reglas gramaticales, puesto que por ser gramatical se espera que la lista sea aplicable a otros contextos en general.

## 5. Conclusión

En nuestro primer intento de lematización de un texto dialectal catalán, hemos comprobado la eficacia del método con un Diccionario y una Gramática, lo mismo que se hace en el ámbito de enseñanza-aprendizaje de lenguas extranjeras en la escuela tradicional. Los nuevos métodos comunicativos, por otra parte, son más prácticos que teóricos: se enseña y se aprende a través de distintas actividades lingüísticas. Esta tendencia también se presenta en el tratamiento de datos lingüísticos, que consiste en reunir los textos anotados y almacenar informaciones estadísticas para aplicarlas a los nuevos textos.

Los dos métodos tienen sus méritos y desventajas. En el método teórico tradicional se busca la exhaustividad de tratamiento, donde se exige la preparación de un buen Diccionario y de una Gramática infalible. En el método práctico de las últimas tendencias, se busca la mayor aplicabilidad posible con un coste relativamente bajo. Se consideraría satisfactorio que el resultado llegase al nivel de más del 95% de respuestas acertadas. Diríamos que el primero

sería trabajo de lingüista, mientras que el segundo, trabajo de especialista de ingeniería informática.[11]

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[12] 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 BIBLIOGRÁFICAS


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, (*pp. 448-463). Oxford: Oxford University Press


[11] Megyesi (2002) se refiere al mérito del etiquetador estadístico derivado de datos suecos, mientras que según el experimento que ha realizado Loftsson (2008) con los textos islándicos, el etiquetador basado en reglas gramaticales presenta mayor eficacia que otros programas de carácter estadístico. Samuelsson y Voutilainen (1997) informan que el programa basado en reglas presenta menos errores que el basado en estadística. Chrupała (2006) propone la combinación de los dos métodos, lo cual está realizado en el proyecto de Guirao y Moreno-Sandoval (2004).

[12] En esta ocasión no hemos tratado la cuestión morfológica, limitándonos a presentar unos ejemplos de reglas sintácticas. Somos conscientes de que es necesario incluir en la Gramática la información morfológica: terminaciones verbales y adjetivales, singular y plural de los sustantivos, sufijos adverbiales, etc. Para la información morfográfica que se toma en consideración, véanse Plissen et al. (2004) y O'Donovan y Troussov (2003).

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/Pillson-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,* (pp. 246-253). Madrid: UNED. 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*, (pp. 145-166). John Benjamins Publishing Company.

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