ESTUDIOS LINGÜÍSTICOS

# LEMATIZACIÓN DE LOS DATOS DE CODEA Y SU UTILIZACIÓN EN ANÁLISIS CUANTITATIVOS SOBRE LA EÑE Y LA HACHE MUDA

LEMMATIZATION OF CODEA DATA AND ITS USE IN QUANTITATIVE ANALYZES ON THE *EÑE* AND THE SILENT *HACHE*

Hiroto Ueda
*University of Tokyo*
uedahiroto@jcom.home.ne.jp
ORCID:0000-0003-3204-609x

Recibido: 07.09.19
Aceptado: 17.10.19
Publicado: 29-12-2019

Resumen

En este artículo explicaremos un método de lematización de los documentos antiguos españoles utilizando los datos de «CODEA» *Corpus de Documentos Españoles Anteriores a 1800* (Sánchez-Prieto et al., 2009) y la herramienta de análisis «LYNEAL» (*Letras y Números en Análisis Lingüísticos*). Nuestro objetivo es presentar el método más sencillo posible de lematización y fácil de realizar con alto grado de precisión. Seguidamente, expondremos dos ejemplos de su utilización en el estudio histórico de la ortografía española: sobre la eñe y la hache muda.

Palabras clave: lematización, documentos antiguos españoles, eñe, hache muda.

Abstract

In this article we will explain a method of lemmatization of Spanish old documents using the data of «CODEA» *Corpus de Documentos Españoles Anteriores a 1800* (Sánchez-Prieto et al., 2009) and the analysis tool «LYNEAL» (*Letras y Números en Análisis Lingüísticos*). Our goal is to present the simplest possible method of lemmatization, easy to perform with high degree of accuracy. Next, we will expose two examples of its use in the historical study of Spanish spelling: on the eñe and the silent hache.

Keywords: lemmatization, Spanish old documents, *eñe*, silent *hache*.

ISSN 1132-0265
http://dx.doi.org/10.12795/PH.2019.v33.i01.10

## 1. Introducción[1]

Se considera que la lematización es el proceso de agrupar "las variantes flexivas y/o derivativas de una palabra a una única forma" (Gómez Díaz 2005: 118), que es un lema o palabra representativa que encabeza un diccionario. En este trabajo, sin embargo, nos concentraremos únicamente en las variantes flexivas, por ej. *voy, vas, va, ..., iré, irás, ..., vaya, vayas, ir, ido, yendo,* etc. → «ir», a exclusión de las derivativas, por ej. *educación, educado, educar* → «educar», puesto que creemos conveniente tratar los lemas derivativos en una etapa posterior a los lemas flexivos, por reunir varios lemas de distintas categorías gramaticales.

La lematización es necesaria y fundamental a la hora de realizar análisis de datos digitales dentro de un determinado corpus lingüístico. Su posibilidad de uso es inmensa en los estudios morfológicos cualitativos y cuantitativos, aplicados a la sociolingüística, geolingüística, psicolingüística, estilística e historia de la lengua, que tratan variaciones y cambios de las unidades léxicas y su constitución morfológica.

A continuación, explicaremos un método de lematización de los documentos antiguos españoles utilizando los documentos de «CODEA» *Corpus de Documentos Españoles Anteriores a 1800* (Sánchez-Prieto et. al. 2009)[2] emitidos en las provincias de Castilla la Vieja (Ávila, Burgos, Logroño, Palencia, Santander, Segovia, Soria, Valladolid) y la herramienta de análisis «LYNEAL» (*Letras y Números en Análisis Lingüísticos*) de elaboración propia[3]. Nuestro objetivo actual es presentar el método más sencillo posible de lematización, fácil de realizar con alto grado de precisión. Seguidamente, expondremos dos ejemplos de su utilización en el estudio histórico de la ortografía española: sobre la eñe y la hache muda.

## 2. Método de lematización

### 2. 1. Unión y separación

Para llevar a cabo la agrupación de las formas léxicas por lema, que corresponde aproximadamente a la entrada del *Diccionario de la lengua española* de la Real Academia Española (DLE)[4], hemos utilizado los dos textos paralelos del corpus: la trascripción paleográfica (TP: transcripción neta del texto con abreviatura entre corchetes angulares) y la presentación crítica (PC: formas desarrolladas de abreviatura, con unión y separación de palabras) de CODEA. Para lematizar las formas paleográficas hemos preparado una edición moderna (EM: con ortografía actual de la

[1] Agradecemos a Leyre Martín Aizpuru, Antonio Moreno Sandoval y Pedro Sánchez de Borja la ayuda prestada durante la preparación de este trabajo. Este estudio se ha llevado a cabo dentro del proyecto "Corpus de Documentos Españoles Anteriores a 1900 (CODEA+2020)", financiado por el Ministerio de Economía (FFI2017-82770-P) y el proyecto "Cronología relativa de los documentos antiguos españoles", financiado por el JSPS KAKENHI (Grant Number 16K02657).

[2] Sánchez-Prieto (2009). El corpus digital se encuentra en http://corpuscodea.es/ [20 de junio de 2019]

[3] http://shimoda.lllf.uam.es/ueda/lyneal/ [20 de junio de 2019]

[4] http://dle.rae.es/ [24 de junio de 2018)]

RAE) a partir de la Presentación Crítica (PC) siguiendo la ortografía académica actual (Real Academia Española 2010).

En el trabajo de hacer corresponder una por una las cuatro formas (TP, PC, EM, Lema), hemos utilizado tres signos: el signo más (+) para unir las formas separadas, el signo igual (=) para unir formas divididas en las dos líneas, de arriba y de abajo, el signo menos, guion (-), para separar las formas unidas. Veamos los ejemplos:

(1) Unión por el signo de más (+)
*TP: ferra<n>do ortega me dixo q<ue> desde la feria de me+djna escriujera al sen<n>or marq<ue>s* (*20 1450 1r 5*)
*PC: Ferrando Ortega me dixo que desde la feria de Medina escriviera al señor marqués,* (*20 1450 1r 5*)

(2) Cambio de línea por el signo de igualdad (=)
*TP: q<ue> enbiase ma<n>dar A j<oa>n ma<n>rriq<ue> q<ue> pre<n>diese A ferrand<o> malo plazeme d<e>- -lo ma<n>=dar* (*20 1450 1r 22*)
*PC: que embiase mandar a Joán Manrique que prendiese a Ferrando Malo, plázeme de lo mandar;* (*20 1450 1r 22*)

(3) Separación por guion (-)
*TP: don<n>a YOLANT mi mugier & con n<uest>ros fijos el Jnffante don Sancho fijo mayor & con don Pedro & don Joh<a>n & don Jaymes. Por grand sabor q<ue> auemos de fazer bien & merçed al Conçeio de Guadalffaiara tan bien a-los de-la villa* (*1277 BU 1r 2*)
*PC: doña Yolant mi mugier, e con nuestros fijos el infante don Sancho, fijo mayor, e con don Pedro, e don Joán e don Jaimes, por grand sabor que avemos de fazer bien e merced al concejo de Guadalfajara, tan bien a los de la villa* (*1277 BU 1r 2*)

De esta manera, por ejemplo, la TP *me+djna* representa la forma unida de la TP original separada: *me djna*. La TP unida con el signo de igual. *ma<n>=dar*, representa *ma<n>* al final de la línea y *dar* al principio de la línea siguiente. La TP *a-los de-la villa* representa las TP originales unidas: *alos dela villa*.

### 2. 2. Agrupación

Hemos hecho corresponder las formas críticas a las formas modernas cuando las dos coinciden tanto en su raíz como en prefijos, sufijos y flexiones. Por ejemplo, la forma verbal *toviéredes* corresponde a la forma moderna *tuviereis*, que pertenece al lema «tener»[5]. A continuación, presentamos la lista de la categoría gramatical (CG) (tabla 1).

Las formas plurales del sustantivo están bajo el mismo lema del singular y las formas femeninas de persona figuran en el mismo lema de la forma masculina, por ejemplo, ***abuelo**, abuelos, abuela, abuelas; **abad**, abadesa; **príncipe**, princesa; **rey**, reina*. El lema de un verbo en forma de infinitivo agrupa todas las formas conjugadas, incluyendo los gerundios y los participios pasados, masculino y femenino,

[5] De aquí en adelante, empleamos las comillas «...» para indicar lemas.

ISSN 1132-0265
http://dx.doi.org/10.12795/PH.2019.v33.i01.10

Tabla 1. Abreviatura, categoría gramatical y ejemplos

| c | Categoría Gramatical | |
|---|---|---|
| adj. | adjetivo | *blanco, blancos, blanca, blancas* |
| adv. | adverbio | *abajo, adelante, antiguamente* |
| art. | artículo | *el, los, la, las, lo, l; un, unos, una, unas* |
| conj. | conjunción | *y; o; maguer; pero* |
| conj. relat. | conjunción / relativo | *que* |
| demos. | demostrativo | *este, estos, esta, estas, esto* |
| indef. | indefinido | *otro, otros, otra, otras* |
| interj. | interjección | *amen* |
| interrog. | interrogativo | *cómo; dónde; qué; quién, quiénes* |
| n. prop. | nombre propio | *Fernando, López* |
| num. | numeral | *uno, dos, tres, cuatro* |
| poses. | posesivo | *mi, mis, mío, míos,* |
| prep. | preposición | *a, con, de, en, por* |
| pron. clit. | pronombre clítico | *me; nos; os; lo, los, la, las, le, se* [*dat.*], *se* [*ref.*] |
| pron. suj. | pronombre de sujeto | *yo; nosotros, nosotras, nos; vosotros, vos; él, ellos, ella, ellas, ello* |
| pron. prep. | pronombre de preposición | *mí, -migo; sí, -sigo; nos, -nosco; vos, -vusco* |
| relat. | relativo | *adonde; cual, cuales; cuanto, cuantos, cuanta, cuantas, cuan; cuyo, cuyos, cuya, cuyas; donde, quien, quienes.* |
| sus. | sustantivo | *abad, abades, abadesa, abadesas; abadía; abuelo, abuelos, abuela, abuelas; rey, reina.* |
| vb. | verbo | *confirmar, confirmo, confirma* |

singular y plural: ***volver,*** *volvería, volvían, volviendo, volviere, volviesen, volvió, vuelta, vuelto, vueltos, vuelva, vuelvan, vuelven.*

### 2. 3. Desambiguación

(1) Palabras vacías

Para la recuperación de información, suele prepararse una lista de palabras vacías en forma de *stop-words* que se excluyen en la búsqueda informática, por ejemplo, artículos, pronombres, preposiciones, conjunciones, verbos auxiliares (*ser, estar, haber*, etc.) y otras palabras de alta frecuencia[6]. La exclusión de las palabras vacías

[6] Buckley et al. (1995), Savoy (1999: 4), Gómez Díaz (2005: 184-185).[6] De aquí en adelante, empleamos las comillas «...» para indicar lemas.

ISSN 1132-0265
http://dx.doi.org/10.12795/PH.2019.v33.i01.10

resulta eficiente puesto que "mejora la precisión y exhaustividad en la recuperación (de información)" (Gómez Díaz 2005: 185).

De nuestra parte, sin embargo, defendemos la inclusión de las palabras vacías precisamente por su alta frecuencia de uso, que no se puede ignorar a la hora de considerar su importancia numérica en los documentos antiguos. A pesar de ser vacías desde el punto de vista de la información, estas palabras son sumamente importantes en el plano lingüístico: «de» (prep.), «y» (conj.), «el» (art.), «que» (conj. relat.), «en» (prep.), «decir» (vb.), «a» (prep.), «lo» (pron.clit.), «por» (prep.), «ser» (vb.), «este» (demos.), «hacer» (vb.), «su» (poses.), «haber» (vb.), «él» (pron. suj.), «todo» (indef.), «mi» (poses.), «con» (prep.), «o» (conj.), «don» (adj.), «no» (adv.), «dar» (vb.), en orden de frecuencia.

(2) Formas ambiguas mayores

Dentro de las palabras vacías de alta frecuencia, destacamos «el» (art.), «lo» (pron. clit.) y «que» (conj. relat.), por el problema que presentan en la lematización. El trabajo de la distinción entre artículo y pronombre clítico en formas de *lo, los, la, las* es relativamente menor, es decir, fácil de realizar, mientras que la distinción de *que* entre conjunción y relativo es sumamente difícil.

Nuestro proceso de lematización ha devuelto el siguiente cálculo:

Tabla 2. Formas ambiguas de artículo definido y pronombre átono

| Forma | F1: Lema | F1: C.g. | F1: Frec. | F2: Lema | F2: C.g. | F2: Frec. |
|---|---|---|---|---|---|---|
| *lo* | «lo» | pron. clit. | 1 419 | «el» | art. | 1 056 |
| *los* | «lo» | pron. clit. | 341 | «el» | art. | 3 729 |
| *la* | «lo» | pron. clit. | 757 | «el» | art. | 6 211 |
| *las* | «lo» | pron. clit. | 196 | «el» | art. | 2 038 |

Según esta tabla, sabemos a ciencia cierta que las formas del artículo son casi siempre más frecuentes que las del pronombre átono con gran diferencia. La excepción es la forma *lo*, en que el pronombre átono es más frecuente que el artículo neutro, con diferencia reducida.

Ante esta situación cuantitativa, pensamos que es mejor tratar *de facto* las formas *lo, los, la, las* como artículo, y posteriormente, cuando estas formas vienen delante de verbos conjugados o detrás de infinitivo, gerundio y formas conjugadas con el signo de separación (-), la regla de categorización las convierte en pronombres: *lo, los, la, las* (art.) → *lo, los, la, las* (pron. clit.). Esto ahorra el trabajo de diferenciar cada caso entre artículo y pronombre desde el principio.

En cuanto a la forma de *que* con su ambivalencia funcional entre conjunción y relativo, gracias al trabajo de Ávila Muñoz (1999: 297) sabemos que su uso más frecuente es como conjunción y no como relativo. Como la distinción de ambas

ISSN 1132-0265
http://dx.doi.org/10.12795/PH.2019.v33.i01.10

categorías gramaticales exige grandes datos de contextos categorizados, de momento la dejamos como una categoría ambigua (conj. / rel.). Sumando las frecuencias de homógrafos de *lo, los, la, las* (15 747) y la de *que* (conj./rel.: 11 368) llega a la cifra de 27 115, que ocupa la proporción de .095 (9.5 %) dentro de la totalidad de 284 984 palabras recogidas en el corpus.

(3) Formas ambiguas menores

A continuación, hemos encontrado otras formas de categoría gramatical ambigua: *bien* (frecuencia: 622), *mando* (369), *era* (300), *poder* (242), *salvo* (123), *fuera* (114), *ruego* (111), *paga* (95), *libre* (90), *ruego* (89), *sí* (84), *juro* (74), *vino* (72), *quito* (50), *demanda* (38), *yantar* (26), *cabe* (22), *deseo* (12), *mora* (8), *prenda* (8), *moro* (7), *toma* (7), *cobre* (6), *manifiesto* (6), *ama* (3), *contralla* (3), *cuesta* (3), *paso* (3), *amo* (2), *armada* (2), *busca* (2), *inserta* (2), *saca* (2), que son 2 830 formas en total (1.0%).

En primer lugar, de esta lista extraemos las formas verbales coincidentes con las formas de otras categorías gramaticales (C. g.) (tabla 3).

En la observación de esta tabla, nos damos cuenta de que estas formas ambiguas son mayoritariamente sustantivas deverbales (*mando, poder, ruego, paga,* etc.) o no deverbales (*era, prenda, moro*). También es importante el hecho de que el sesgo de la frecuencia entre las dos formas homógrafas es tan grande que la frecuencia mínima entre las dos resulta bastante reducida, a excepción del caso de *era* (verbo: 124). Por esta razón, conviene asignar el lema y la categoría gramatical de mayor frecuencia a todas estas formas ambiguas, por ejemplo *mando* (vb.), *era* (sus.), *poder* (sus.), etc., para posteriormente aplicar automáticamente las reglas de secuencia de categoría gramatical. En el proceso, destacamos las formas agramaticales por medio de un asterisco ( * ), por ejemplo, *mando* (vb.) detrás de artículo o posesivo:

*mando* (vb.) → (*sus.) / art. pos. __

Por la marca de *sus., indicada por el programa, el analizador se fijará en los contextos para determinar el lema y la categoría correctamente.

La misma operación es aplicable en los dos casos restantes: *bien* (sus. - adv.) y *sí* (pronombre proposicional - adv.) (tabla 4).

La forma *bien* sería *de facto* adverbio (adv.) y la forma *sí* sería pronombre preposicional (pron. prep.) y posteriormente aplicando la regla de secuencia de categoría se corregirán de lema y categoría de manera automática o manual.

ISSN 1132-0265
http://dx.doi.org/10.12795/PH.2019.v33.i01.10

Tabla 3. Formas ambiguas menores

| Forma | F1: Lema | C.g. | Frec. | F2: Lema | C.g. | Frec. | Suma |
|---|---|---|---|---|---|---|---|
| *mando* | «mando» | sus. | 2 | «mandar» | vb. | **367** | 369 |
| *era* | «era ‹tierra'» | sus. | **176** | «ser» | vb. | 124 | 300 |
| *poder* | «poder» | sus. | **227** | «poder» | vb. | 15 | 242 |
| *salvo* | «salvo» | adj. | **122** | «salvar» | vb. | 1 | 123 |
| *fuera* | «fuera» | adv. | **109** | «ser» | vb. | 5 | 114 |
| *ruego* | «ruego» | sus. | **88** | «rogar» | vb. | 23 | 111 |
| *paga* | «paga» | sus. | **92** | «pagar» | vb. | 3 | 95 |
| *libre* | «libre» | adj. | **89** | «librar» | vb. | 1 | 90 |
| *ruego* | «ruego» | sus. | **88** | «ruego» | vb. | 1 | 89 |
| *juro* | «juro» | sus. | **71** | «jurar» | vb. | 3 | 74 |
| *vino* | «vino» | sus. | **66** | «venir» | vb. | 6 | 72 |
| *quito* | «quito» | adj. | **34** | «quitar» | vb. | 16 | 50 |
| *demanda* | «demanda» | sus. | **29** | «demandar» | vb. | 9 | 38 |
| *yantar* | «yantar» | sus. | **25** | «yantar» | vb. | 1 | 26 |
| *cabe* | «cabe» | prep. | 11 | «caber» | vb. | 11 | 22 |
| *deseo* | «deseo» | sus. | 2 | «desear» | vb. | **10** | 12 |
| *mora* | «moro» | sus. | 1 | «morar» | vb. | **7** | 8 |
| *prenda* | «prenda» | sus. | **6** | «prender» | vb. | 2 | 8 |
| *moro* | «moro» | sus. | **5** | «morar» | vb. | 2 | 7 |
| *toma* | «toma» | sus. | 1 | «tomar» | vb. | **6** | 7 |
| *manifiesto* | «manifiesto» | adj. | 1 | «manifestar» | vb. | **5** | 6 |
| *cobre* | «cobre» | sus. | **4** | «cobrar» | vb. | 2 | 6 |
| *ama* | «ama» | sus. | 1 | «amar» | vb. | 2 | 3 |
| *contralla* | «contralla» | sus. | **2** | «contrallar» | vb. | 1 | 3 |
| *cuesta* | «cuesta» | sus. | **2** | «costar» | vb. | 1 | 3 |
| *paso* | «paso» | sus. | **2** | «pasar» | vb. | 1 | 3 |
| *inserta* | «inserto» | adj. | 1 | «insertar» | vb. | 1 | 2 |
| *amo* | «amo» | sus. | 1 | «amar» | vb. | 1 | 2 |
| *busca* | «busca» | sus. | 1 | «buscar» | vb. | 1 | 2 |
| *saca* | «saca» | sus. | 1 | «sacar» | vb. | 1 | 2 |
| *armada* | «armada» | sus. | 1 | «armar» | vb. | 1 | 2 |

ISSN 1132-0265
http://dx.doi.org/10.12795/PH.2019.v33.i01.10

Tabla 4. Frecuencia de *bien* y *sí*

| Forma | F1: C.g.. | Frec.. | F2: C.g.. | Frec.. | Suma | Min(F1,F2) |
|---|---|---|---|---|---|---|
| *bien* | sus. | 228 | adv. | **394** | 622 | 228 |
| *sí* | pron. prep. | **81** | adv. | 3 | 84 | 3 |

## 3. Análisis de datos lematizados

A modo de ejemplo de cómo se utilizan los datos lematizados, escogemos dos temas de la historia de las grafías españolas: la *ñ* en todos los lemas y la ***h*** muda del lema «haber». Ambas cuestiones serían sumamente difíciles de tratar, si no fuera por los datos lematizados.

### 3. 1. Grafema *ñ*

El primer ejemplo del análisis de lemas se trata del grafema ñ. Al respecto, Torrens Álvarez (2018: 175) explica el origen de la eñe en el siguiente párrafo:

> ***ñ:*** letra emblemática del abecedario español, su morfología deriva de la costumbre de abreviar *nn* escribiendo una sola *n* con lineta abreviativa superpuesta, lineta que igual se empleaba en *señor* 'señor' o *año* 'año' que en *cōnde* 'conde' o *cātauā* 'cantaban'. (...) fueron varios los ensayos primitivos que se hicieron para representar /ɲ/, como los etimológicos *ni, ng, gn,* además de *nn*, así como la *n* simple (...), pero el castellano pronto se decidió por la doble *n*, explícita o con una *n* abreviada. Es difícil precisar en qué fecha puede hablarse de *ñ* como tal letra, pues todavía en el siglo XVI su tilde no difiere de la que abrevia otras nasales en otros contextos o, incluso, otras letras distintas.

Para observar los altibajos diacrónicos en la frecuencia de algunas manifestaciones gráficas, hemos buscado los lemas que contienen o han contenido la grafía ***ñ*** a lo largo de la historia, que se cuentan 3 841 en total. Para fijarnos en la parte principal de la variación, hemos seleccionado los primeros doce lemas: «año», «señor», «doña», «conocer», «daño», «viña», «señorío», «señoría», «dueño», «empeñar», «aniversario», «pequeño», en total, 3 477.

> ***<n>n*:** *se<n>nor* (7), *a<n>nos* (2), *co<n>nosçemos* (2), *co<n>noscida* (2), *ui<n>nas* (2), *vi<n>nas* (2), *a<n>niuersario* (1), *co<n>nosçida* (1), *co<n>nuçuda* (1), *do<n>na* (1), *uj<n>nas* (1), en total 22.
> ***nn*:** *annos* (27), *anno* (21), *donna* (13), *cadanno* (5), *danno* (5), *connosçuda* (4), *connosçida* (3), *vinnas* (3), *duennas* (2), *sennor* (2), *anniu<*>sario* (1), *anniu<*>ssario* (1), *connocida* (1), *connoçida* (1), *etc.*, en total 98.

ISSN 1132-0265
http://dx.doi.org/10.12795/PH.2019.v33.i01.10

***n*<*n*>**: *sen*<*n*>*or* (499), *an*<*n*>*o* (418), *an*<*n*>*os* (418), *don*<*n*>*a* (220), *sen*<*n*>*ora* (109), *dan*<*n*>*o* (101), *sen*<*n*>*ores* (67), *sen*<*n*>*or*<***>*s* (61), *sen*<*n*>*orios* (61), *sen*<*n*>*orio* (54), *dan*<*n*>*os* (43), *ssen*<*n*>*or* (43), *etc.*, en total 2 408.
***g*<*n*>**: *cog*<***>*sco* (1)
***gn*: *cognosçemos* (6), *cognosco* (6), en total 12.
***n*:** *dona* (32), *conosco* (26), *conosçemos* (16), *conozco* (11), *anos* (8), *aniu*<***>*ssario* (7), *conosçer* (7), *conosçido* (7), *ano* (6), *ssenorio* (6), *enpenar* (5), *senor*<***>*s* (5), *anju*<***>*sarios* (4), *senora* (4), *vjnas* (4), *anjuersario* (3), *etc.*, en total 286.
<ñ>: *año* (58), *doña* (56), *años* (54), *señor* (39), *señora* (20), *señores* (11), *viñas* 11, *pequeña* (8), *biña* (7), *pequeño* (7), *pequeñas* (6), *daño* (4), *señoria* (4), *dueños* (3), *daños* (2), *dueño* (2), *pequeños* (2), *señorio* (2), *señorios* (2), *etc.*, en total 299.
<***n*>*pn*: *da*<*n*>*pno* (1)
***pn*: *dapnos* (10), *dapno* (1), en total 11.
***pn*<*n*>: *dapn*<*n*>*o* (14), *dapn*<*n*>*os* (8), en total 22.
***ynn*: *aynnos* (1)

A partir de estas formas y sus frecuencias, hemos elaborado la tabla de distribución cronológica[7]:

Tabla 5. Grafías del grafema ***ñ***. Distribución original. Frecuencia absoluta

| ñ | 1200 | 1250 | 1300 | 1350 | 1400 | 1450 | 1500 | 1550 | 1600 | 1650 | 1700 | 1750 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *<n>n* | 6 | 14 | 1 | — | 2 | — | — | — | — | — | — | — |
| *gn* | — | — | — | — | — | 10 | 3 | — | — | — | — | — |
| *n* | 6 | 36 | 33 | 19 | 33 | 31 | 52 | 19 | 8 | 10 | 4 | 4 |
| *n<n>* | 8 | 416 | 329 | 236 | 367 | 520 | 522 | 40 | 2 | — | — | — |
| *nn* | 12 | 80 | 4 | 1 | — | 1 | — | — | — | — | — | — |
| *pn<n>* | — | — | — | — | 15 | 9 | 9 | — | — | — | — | — |
| ñ | — | — | — | — | — | 1 | 9 | 58 | 90 | 29 | 67 | 45 |
| Total | 32 | 546 | 367 | 256 | 417 | 572 | 595 | 117 | 100 | 39 | 71 | 49 |

La tabla anterior ha sido conseguida por medio de nuestra herramienta LYNEAL, que facilita la organización de los datos en dos dimensiones: la forma en el orden alfabético en posición vertical y la cronología horizontal dividida en franjas con

[7] Las grafías de poca frecuencia, *g<n>, <n>pn / pn, ynn,* las sumamos a la forma representante, *gn, pn<n>, nn,* respectivamente.

ISSN 1132-0265
http://dx.doi.org/10.12795/PH.2019.v33.i01.10

intervalos de 50 años, donde cada año representa el inicio de la franja (por ejemplo, 1200 comprende desde 1200 hasta 1249).

La Frecuencia absoluta tiene el mérito de representar la magnitud cuantitativa real de cada caso, a costa del inconveniente de no poder comparar los datos entre ellos. Por ejemplo, la frecuencia 6 de <*n*> en 1200 no manifiesta necesariamente su minoría con respecto a la frecuencia 14 que se observa en 1250, puesto que la totalidad de cada año es diferente (32 y 546). Para solucionar este problema suele utilizarse la frecuencia relativa en forma de porcentaje (%):

Tabla 6. Grafías del grafema ñ. Distribución original. Porcentaje (%)

| ñ | 1200 | 1250 | 1300 | 1350 | 1400 | 1450 | 1500 | 1550 | 1600 | 1650 | 1700 | 1750 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| <*n*>*n* | 18.8 | 2.6 | 0.3 | | 0.5 | — | — | — | — | — | — | — |
| *gn* | — | — | — | — | — | 1.7 | 0.5 | — | — | — | — | — |
| *n* | 18.8 | 6.6 | 9.0 | 7.4 | 7.9 | 5.4 | 8.7 | 16.2 | 8.0 | 25.6 | 5.6 | 8.2 |
| *n*<*n*> | 25.0 | 76.2 | 89.6 | 92.2 | 88.0 | 90.9 | 87.7 | 34.2 | 2.0 | — | — | — |
| *nn* | 37.5 | 14.7 | 1.1 | 0.4 | — | 0.2 | — | — | — | — | — | — |
| *pn*<*n*> | — | — | — | — | 3.6 | 1.6 | 1.5 | — | — | — | — | — |
| *ñ* | — | — | — | — | — | 0.2 | 1.5 | 49.6 | 90.0 | 74.4 | 94.4 | 91.8 |
| Total | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |

A nuestro modo de ver, la frecuencia relativa causa el problema del desajuste con la realidad. Por ejemplo, el porcentaje de *nn* en 1200 (37.5 %, 12 / 32 = .372) resulta ser excesivo en contraste con la cifra de la misma grafía en 1250 (14.7 %, 80 / 546 = .147) que es una cantidad poco significativa. Este problema nace de una expectación excesiva en el sentido de que 12 dentro de 32 se espera que llegaría a 37.5 dentro de 100 (37.5 %). Por ejemplo, nadie piensa que 4 éxitos conseguidos en 10 ensayos garanticen en futuro 40 éxitos en 100 ensayos (40 %), que es una expectación excesiva, demasiado oportunista, desde el punto de vista probabilístico. De esta manera, las frecuencias relativizadas no reflejan necesariamente la realidad estadística, lo que causa un serio problema al analizador, quien a veces no sabe cómo tratar los porcentajes.

Para dar solución a los problemas que causan tanto la frecuencia absoluta, incomparable, como la frecuencia relativa, excesivamente evaluada, conviene realizar la conversión de la tabla de frecuencia absoluta en forma de distribución diagonalizada (Ueda 2017). Para obtener una visión de tendencia histórica, hemos llevado a cabo una operación de 'diagonalización' que consiste en cambiar el orden vertical de las formas para que la distribución de la frecuencia se ponga concentrada en la línea diagonal, que parta del punto superior izquierdo y termine en el punto inferior derecho. De esta manera observamos cada frecuencia real situada dentro del

ISSN 1132-0265
http://dx.doi.org/10.12795/PH.2019.v33.i01.10

Tabla 7. Grafías del grafema ñ. Distribución diagonalizada. Frecuencia absoluta.

| ñ | 1200 | 1250 | 1300 | 1350 | 1400 | 1450 | 1500 | **1550** | 1600 | 1650 | 1700 | 1750 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *nn* | 12 | 80 | 4 | 1 | | 1 | | | | | | |
| *<n>n* | 6 | 14 | 1 | | 2 | | | | | | | |
| *n<n>* | 8 | 416 | 329 | 236 | 367 | 520 | 522 | **40** | 2 | | | |
| *n* | 6 | 36 | 33 | 19 | 33 | 31 | 52 | **19** | 8 | 10 | 4 | 4 |
| *pn<n>* | | | | | 15 | 9 | 9 | | | | | |
| *gn* | | | | | | 10 | 3 | | | | | |
| ñ | | | | | | 1 | 9 | **58** | 90 | 29 | 67 | 45 |

contexto graduado global, lo que permite la evaluación de cada cantidad de manera convenientemente comparable y correctamente evaluable (tabla 7).

Por medio de la distribución diagonalizada de las frecuencias, que obtenemos por reordenación vertical de las formas lingüísticas, podemos fijarnos en las formas con relación a los años de cada porción de la línea diagonal. Por ejemplo, *nn* y *<n>* corresponden a las franjas de 1200 y 1250, mientras que *n<n>* es relativamente posterior, de 1250 a 1500. La simple *n* es omnipresente con mayor concentración desde 1250 hasta 1550. Las formas de *pn<n>* se presentan en 1400 con cierta prolongación posterior en 1450 y 1500. La grafía *gn* es peculiar en 1450. Finalmente, la eñe (*ñ*) empieza a predominar en 1550. Es importante notar que esta descripción es una cronología gradual en orden de 1200 a 1750, lo que se consigue gracias al proceso de diagonalización, realizada por un programa informático instalado en nuestro sistema.

En esta tabla, no se trata de comparar las frecuencias de manera independiente con respecto a otras, sino de situarlas en la visualización general de distribución. Por esta tabla de distribución cronológica concentrada en las zonas diagonales, podemos observar la transición histórica solapada de frecuencias de *nn* → *<n>n* → *n<n>* → *n* → *pn<n>* → *gn* → *ñ*, que no contradice la explicación de Torrens Álvarez (loc. cit.).

De nuestra parte, hemos incluido una doble *n* y dos variantes de *<n>* abreviada: *nn*, *<n>n* y *n<n>* en este orden: *nn* → *<n>n* → *n<n>,* lo cual parece indicar que el origen de eñe se encuentra en *nn*, que coincide con *<n>n*, minoritaria, en la posición implosiva de *<n>* abreviada, y *n<n>*, predominante, con una mayor extensión cronológica posterior. La simple *n* es una forma de base en competencia con las restantes marcadas. Las variantes *pn<n>* y *gn* son peculiares de *daño* y *conocer*, respectivamente. Finalmente llegamos a la grafía actual *ñ*, con supremacía numérica sobre las otras variantes *n<n>* y *n* en 1550, año en que reconocemos el inicio predominante de ella.

### 3. 2. H muda

A pesar de que en el español actual la grafía *h* muda se escribe explícitamente, se han venido registrando distintas formas sin ella a lo largo de siglos. Salvador y Lodares (2001: 124) explican su historia de la siguiente manera:

> (...) en las escrituras más antiguas [la *h*] puede muy bien no aparecer: *omne, onor, onra, oz* (hoz), *auer* (haber) se localizan con suma facilidad en documentos antiguos. Pero, al fin y al cabo, el español es hijo del latín, y en periodos en los que se ha recurrido a la madre para enriquecer al hijo se ha reconstruido con fidelidad la ortografía latina. Esta reconstrucción de haches es visible desde mediados del siglo XIII, con Alfonso X, y ya no se detendrá jamás: sigue con los autores del siglo XV enamorados del latín; con Nebrija después (...); luego vino Francisco de Robles, sigue con Sebastián de Covarrubias más tarde y, en fin, se consolida con la Academia de modo que las palabras que en latín llevaban *h* pasarán a escribirse con ella en español.

En contraste, Torrens Álvarez (2018: 173) presenta una visión diferente: "(...) lo general es que en las voces patrimoniales la *h-* inicial se pierde, con lo que a lo largo de toda la Edad Media y los Siglos de Oro lo normal es escribir *aver* 'haber', *omne* 'hombre', *estoria* 'historia'". Efectivamente los antiguos documentos notariales presentan la situación histórica de manera un tanto distinta de la explicación de Salvador y Lodares (loc. cit.). En primer lugar, la *h* no es que *pueda* no aparecer, sino casi siempre *no aparece* en los siglos medievales como veremos más adelante en los paradigmas verbales de «haber» y en las formas del sustantivo «hombre».

Las excepciones son las formas conjugadas de *he, ha* y *han*, que fueron bastante frecuentes a lo largo de historia, inclusive en la Edad Media. Creemos que estas frecuencias anómalas no se deben a Alfonso X (Salvador y Lodares loc. cit.), puesto que la *h* aparece también antes de 1250 en las tres formas conjugadas. Por otra parte, los innumerables casos de no aparición de *h* en otras formas verbales en la franja de 1250 serían contradictorios con la supuesta reconstrucción alfonsí. Veamos los datos de distintas formas de los lemas «haber» y «hombre»:

Tabla 8. Frecuencia absoluta de «haber», 'he': *e, h<e>, he*

| 'he' | 1200 | 1250 | 1300 | 1350 | 1400 | 1450 | 1500 | 1550 | 1600 | 1650 | 1700 | 1750 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *e* | | 2 | 6 | 4 | | | | | | | | |
| *h<e>* | 1 | 32 | 10 | 7 | 12 | 34 | 20 | | 1 | | | 4 |
| *he* | | 7 | 3 | | | 2 | 1 | 1 | 4 | 2 | | |

Tabla 9. Frecuencia absoluta de «haber», «*ha*»: *a*, *ha*.

| 'ha' | 1200 | 1250 | 1300 | 1350 | 1400 | 1450 | 1500 | 1550 | 1600 | 1650 | 1700 | 1750 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *a* | 2 | 12 | 18 | 12 | 3 | 3 | 20 | 9 | 32 | 6 | 6 | 1 |
| *ha* | 1 | 1 | 11 | 12 | 14 | 40 | 55 | 4 | 12 | 3 | | 8 |

ISSN 1132-0265
http://dx.doi.org/10.12795/PH.2019.v33.i01.10

Tabla 10. Frecuencia absoluta de «haber», 'haber': *auer, haver, aber; hauer, haber, haber*

| 'haber' | 1200 | 1250 | 1300 | 1350 | 1400 | 1450 | 1500 | 1550 | 1600 | 1650 | 1700 | 1750 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *au<er>* | | | 2 | 14 | 1 | | 2 | | | | | |
| *auer* | 9 | 45 | 24 | 15 | 24 | 7 | 4 | 10 | 13 | 7 | | |
| *aue<r>* | | | 1 | | 4 | 1 | | | | | | |
| *av<er>* | | | | | | 33 | 29 | | | | | |
| *aver* | | | | | 4 | 11 | 22 | 4 | 1 | | | |
| *aber* | | | 1 | | | | 1 | 3 | 1 | | 1 | 1 |
| *hauer* | | | | | | | 1 | 1 | 8 | 2 | 1 | 3 |
| *haber* | | | | | | | | | 2 | | | 2 |
| *haver* | | | | | | | | | | | | 2 |

Tabla 11. Frecuencia absoluta de «hombre»: omne, ome, ombre; hombre

| «hombre» | 1200 | 1250 | 1300 | 1350 | 1400 | 1450 | 1500 | 1550 | 1600 | 1650 | 1700 | 1750 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *omne* | 3 | 79 | 58 | 26 | 44 | 49 | 9 | | | | | |
| *ome* | | 5 | 6 | 2 | 7 | 4 | 11 | 1 | | | | 36 |
| *ombre* | | 2 | 3 | 1 | | 3 | 4 | | | | 1 | 14 |
| *hombre* | | | | | | | 5 | 4 | 15 | 1 | 1 | 1 |

La misma tendencia de no escribir la *h* en los siglos medievales antes de 1500, se confirma en los datos de los lemas «honor», «honra», «honrar», etc. En este sentido, Sánchez-Prieto (1998: 119) menciona estas dos posibilidades en la forma de *ha.* El mismo autor en su trabajo posterior (2004: 436-437) reconoce mayor peso en "el aumento de la imagen visual".

De nuestra parte, sin negar la función de *h* como "incremento del contorno gráfico" (Sánchez-Prieto, 1998: loc. cit.) o la extensión analógica de la *h* de *ha* a la forma de *he* del mismo verbo «haber»[8], ambas como factores coadyuvantes, proponemos extender la idea del "valor diacrítico" (ibid.) para explicar no solamente la forma *ha* y *ha<n>* con respecto a la preposición «a», sino también *he* del verbo «haber» frente a la conjunción «y» en forma de *e*. A pesar de que en la Edad Media la inmensa mayoría de representación gráfica de la conjunción «y» se realizaba por medio del signo tironiano &, se observa cierta cantidad de otras variantes importantes (*et, e*) por lo que se considera el tironiano como alógrafo de *e*(*t*) antes de 1500. De ahí que naciera la función distintiva de *h* en *he* del verbo «haber», incluso

[8] Torrens Álvarez (comunicación personal).

ISSN 1132-0265
http://dx.doi.org/10.12795/PH.2019.v33.i01.10

con la grafía tironiana predominante, es decir, el escribiente escogería la forma *he* en «haber» para distinguirla de la forma *e*(*t*), aun cuando escribiera en su lugar la forma tironiana[9].

Todas las formas de la misma conjunción pertenecerían al mismo lema «y» sin distinción cronológica clara. En la siguiente tabla observamos la situación histórica con graduación tendente desde *et*, pasando sucesivamente por *&, e, i /j*, para llegar a *y*, siempre de manera solapada sin división tajante:

Tabla 12. Frecuencia de las formas de conjunción «y».

| «y» (conj.) | 1200 | 1250 | 1300 | 1350 | 1400 | 1450 | 1500 | 1550 | 1600 | 1650 | 1700 | 1750 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *et* | 21 | 452 | 395 | 150 | 225 | 24 | 9 | | | | | |
| *&* | 328 | 3402 | 2352 | 1463 | 2275 | 3550 | 2726 | 40 | | | | |
| *e* | 22 | 158 | 106 | 192 | 379 | 626 | 816 | 317 | 9 | 1 | | |
| *i, j* | | 1 | | | | 16 | 16 | | 1 | | 1 | |
| *y* | 1 | 22 | 4 | 3 | 11 | 221 | 750 | 1023 | 551 | 301 | 456 | 268 |

El uso de la *h* muda actual está explicado por el "criterio etimológico" (Real Academia Española 2010, 142). Aun reconociendo la importancia del mismo criterio, creemos también conveniente pensar en la *h* que hemos observado en las tres formas frecuentes del verbo «haber», *he, ha, han*, verbo muy importante por su carácter auxiliar con alto grado de gramaticalización. La alta frecuencia de las tres formas mencionados apoyaría el uso de *h* muda en otras formas posteriores pertenecientes al mismo paradigma verbal, *hemos, habemos, habéis, había, habiendo,* etc., con generalización a todas las palabras sucedidas de la H- latina, tales como *hombre, honor, hábil,* etc. Sin apoyo de estas formas antecedentes, el mencionado criterio etimológico no hubiera funcionado tan fácilmente con la unificación casi completa.

## 4. Conclusiones

En el mundo de la lingüística de corpus, se discute sobre la necesidad de lematizar los textos objeto de investigación lingüística. En este sentido, Sinclair (1991: 21-22) sostenía la idea de mantener los textos limpios, no lematizados, para evitar la mezcla de anotaciones impuestas por la lematización con sus propios criterios

[9] Marcet Rodríguez (2010: 66) trata la fórmula notarial *he e deuo auer,* donde se observa el uso de *h*, como "un buen ejemplo de la *variatio* gráfica como recurso estilístico, aunque más bien parece tratarse de un intento de evitar la confusión de la primera persona singular del verbo *haber* con la conjunción copulativa *e*". Esta observación es importante y proponemos considerar el mismo intento de evitar la confusión con la conjunción «e», no solo en este contexto particular sino también fuera de él, en general.

ISSN 1132-0265
http://dx.doi.org/10.12795/PH.2019.v33.i01.10

gramaticales, lo que causaría un problema a la hora de analizarlos con los objetivos y métodos diferentes de investigadores. En cambio, McEnery & Hardie (2014: 153-162) se oponen a la opinión de Sinclair alegando que el marco teórico preexistente puede ayudar al lingüista de la misma manera que la intuición preconcebida suya. Ante este debate, pensamos que, aparte de la posibilidad de preparar múltiples versiones del texto, limpio y con distintos niveles de anotación, la solución se encuentra en la mínima anotación posible que permita su amplia utilización posterior. Concretamente nuestra anotación se limita solo al lema y parte de la oración a exclusión de análisis morfosintácticos profundos. De esta manera, creemos que se facilita la búsqueda general con lema y parte de oración, que abre un amplio campo de aplicación.

Con respecto al método concreto de lematización, hemos propuesto realizar un análisis estadístico previo del texto en general. Nuestro análisis previo muestra que las listas necesarias para la lematización son enormes pero repetitivas en un corpus de características homogéneas, en nuesro caso, en documentos notariales. Actualmente el ordenador, gracias a los últimos desarrollos tecnológicos, permite un almacenamiento grande de datos de acceso inmediato, lo que facilita las distintas operaciones de lematización de manera fácil y eficiente.

Antes intentábamos elaborar un conjunto de reglas sin saber si eran necesarias en realidad. Por ejemplo, la forma *trabajo* puede ser no solamente un sustantivo sino también la primera persona singular del presente de indicativo del verbo «trabajar». No obstante, en realidad la frecuencia del sustantivo suele ser más alta que la forma conjugada del verbo. Por esta razón, pensamos que es mejor tratarlo como sustantivo *a priori* y cuando la misma forma aparece en el contexto de verbo, en ese momento lo averiguamos *in situ* en la línea de concordancia y lo corregimos como verbo si es necesario. De esta manera, nuestro método se diferencia de los métodos automáticos habituales que seleccionan los candidatos de parte de oración y el lema[10] en el sentido de que en el primer momento no hacemos la selección ni el análisis sino admitimos la forma entera dotada de un único lema y única parte de la oración, y posteriormente, si es necesario, procedemos al reanálisis.

En la historia de la lingüística moderna del siglo pasado, se presentaron dos modelos del análisis morfológico: IA (*item and arrangement*), que consiste en preparar la lista de morfemas y su modo de combinación, e IP (*item and process*), que explica la forma flexiva a partir de la forma inicial con la aplicación sucesiva de reglas de proceso (Hockett 1954). Nuestro método no pertenece a ninguno de los dos sino que se parece más al modelo tradicional de WP (*word and paradigm*) (Hockett *ibid*. 386), que nos recuerda el antiguo método de enseñanza de lenguas extranjeras: presentar la forma representativa (lema) junto con todas sus formas flexivas al lado. De todos modos, nuestro método sintético difiere del modelo WP en que introducimos el concepto de la probabilidad sesgada. Los tres métodos presentan sus

[10] Para la explicación detallada de varios modelos automáticos aplicados al español, véase el trabajo reciente de Moreno Sandoval (2019, 145-170).

ISSN 1132-0265
http://dx.doi.org/10.12795/PH.2019.v33.i01.10

beneficios e inconvenientes según el caso y ninguno es mejor que el otro en general. Nuestra decisión de optar por WP obedece a las razones prácticas: la distribución sesgada de unidades lingüísticas y el avance tecnológico actual que permite el uso de una memoria grande.

El sesgo cuantitativo que se observa en distintos aspectos de la lengua es tan considerable que lo necesitamos tener muy en cuenta tanto en el procesamiento de datos textuales como en los análisis lingüísticos de textos. La mayoría de las veces, el conjunto de elementos en oposición se divide en dos o más miembros de manera bastante desigual. Dentro del conjunto sesgado de elementos, nos interesa, especialmente, la mayoría cuantitativa de casos, que determina la tendencia general de la historia, como hemos visto en la sección 3.

Al respecto, creemos conveniente citar un párrafo de Halliday (1991: 33-34), que consideramos sugerente e importante:

> Diachronically, frequency patterns as revealed in corpus studies provide explanations for historical change, in that when interpreted as probabilities they show how each instance both maintains and perturbs the system. 'System' and 'instance' are of course not different things; they form yet another complementarity. (…) To the 'instance' observer, the system is the potential, with its set of probabilities attached; each instance is by itself unpredictable, but the system appears constant through time. To the 'system' observer, each instance redefines the system, however infinitesimally, maintaining its present state or shifting its probabilities in one direction or the other (…) it is the system-observer who perceives depth in time; (…)

Sobre la opinión de Halliday, Stubbs (2007: 138) comenta: "Ferquency is a fact about past events, but probability is a prediction about future events". A nuestro entender, la frecuencia y la probabilidad son dos aspectos de la misma cuantificación del fenómeno. La frecuencia sin más se refiere a la frecuencia absoluta, de la que se deduce la frecuencia relativa calculada en base a la totalidad. La estadística enseña que la probabilidad es la misma que la frecuencia relativa, que posee el rango de 0 (probabilidad nula) a 1 (seguridad absoluta). Naturalmente, la probabilidad puede servir para la predicción sobre eventos del futuro, pero también es útil para una evaluación del pasado o del presente por medio de relativización y generalización cuantitativa. Sin embargo, ambos tipos de frecuencia, absoluta y relativa, tienen inconvenientes graves: imposibilidad de comparar y desajuste con la realidad o expectación excesiva, respectivamente. Hemos propuesto una solución en forma de distribución diagonalizada (sección 3. 1).

Seguidamente, Hallyday (loc. cit.) continúa:

> but the transformation of instance into system can be observed only through the technology of the corpus, which allows us to accumulate instances and monitor the diachronic variation in their patterns of frequency.

ISSN 1132-0265
http://dx.doi.org/10.12795/PH.2019.v33.i01.10

Para "acumular casos y monitorizar la variación diacrónica en su patrón de frecuencia" (Hallyday, *loc. cit.*), creemos haber demostrado la utilidad de los datos lematizados almacenados en una herramienta que facilita la búsqueda exclusiva y exhaustiva y la visualización en forma de tablas adecuadamente ordenadas.

## Referencias bibliográficas


Ávila Muñoz, A. (1999). *Léxico de frecuencia del español hablado en la Ciudad de Málaga*. Málaga, España: Universidad de Málaga.

Buckley, C., Salton, G., Allen, J., y Singhal, A. (1995). Automatic query expansion using SMART. *Proceedings of the TREC'3 Conference,* 69-80. Gaithersburg, MA: NIST publication.

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

Halliday, M. A. K. (1991). Corpus studies and probabilistic grammar. En Aijmer y B. Altenberg (Eds.), *English corpus linguistics. Studies in honour of Jan Svartvick* (pp. 30-43). London, UK: Longman.

Hockett, C. F. (1954). Two models of grammatical description. *Word,* 10, 210-231. https://doi.org/10.1080/00437956.1954.11659524

Marcet Rodríguez, V. J. (2010). De nuevo sobre los usos y valores de la grafía h en la escritura medieval leonesa. En M. T. Encinas Manterola et al. (Eds.). *Ars longa. Diez años de Asociación de Jóvenes Investigadores de Historiografía e Historia de la Lengua Española* (pp. 63-80). Salamanca, España: Universidad de Salamanca.

McEnery, T. & Hardie, A. (2012). *Corpus linguistics.* Cambridge, UK: Cambridge University Press. https://doi.org/10.1093/oxfordhb/9780199276349.013.0024

Moreno Sandoval, A. (2019). *Lenguas y computación.* Madrid, España: Editorial Síntesis.

Real Academia Española. (2010). *Ortografía de la lengua española*. Madrid, España: Espasa Libros.

Salvador, G. y Lodares, J. R. (2001). *Historia de las letras.* Madrid, España: Espasa.

Sánchez-Prieto, P., Paredes García, F. R., Martínez Sánchez, Miguel Franco, R. Simón Parra, M. y Vicente Miguel, I. (2009). El Corpus de Documentos Españoles Anteriores a 1700 (CODEA). En A. Enrique-Arias (Ed.), *Diacronía de las lenguas iberorrománicas: Nuevas aportaciones desde la lingüística de corpus* (pp 25-38). Madrid/Frankfurt am Main, España/Alemania: Iberoamericana-Vervuert. https://doi.org/10.31819/9783865278685-003

Savoy, J. (1999). A stemming procedure and stopword list for general French corpora. *Journal of the American Society for Information Science,* 50(10), 944-952. https://doi.org/10.1002/(SICI)1097-4571(1999)50:10<944::AID-ASI9>3.0.CO;2-Q

Sinclair, J. (1991). *Corpus, concordance, collocation.* Oxford, UK: Oxford University Press.

Stubbs, M. (2007). On texts, corpora and models of language. En Hoey, E., Mahlberg, M., y Teubert, W (Eds.), *Text, discourse and corpora. Theory and analysis* (pp. 127-162). New York, EEUU: Continuum.

Torrens Álvarez, M. J. (2018). *Evolución e historia de la lengua española.* 2a edición. Madrid, España: Arco / Libros.

Ueda, H. (2017). Unilateral correspondence analysis applied to Spanish linguistic data in time and space. *Sixteenth International Conference on Methods in Dialectology.* National Institute for Japanese Language and Linguistics, Tokyo, 10 August, 2017. https://lecture.ecc.u-tokyo.ac.jp/~cueda/kenkyuchiricorrespondencecorrespondence2017.pdf

____ (2018). Tratamiento lingüístico y matemático de textos digitales españoles. Presentación del Programa LEXIS-web. *Actas del IX Congreso de la Asociación Asiática de Hispanistas* (*Bangkok, 2016*), 617-630.
http://www.sinoele.org/images/Revistá17/monograficos/AAH_2016/AAH_2016_hiroto_ueda.pdf

ISSN 1132-0265
http://dx.doi.org/10.12795/PH.2019.v33.i01.10