Co-funded by the
Erasmus+ Programme
of the European Union

EULALIA

European Latin Linguistic Assessment

Erasmus+ Strategic Partnership for Higher Education (2019-2022)

(2019-1-IT02-KA203-062286)

https://site.unibo.it/eulalia/en

O1: European Latin Language Certification – Basic Level

Methodological and Pedagogical tools

LEXICON

(Spanish Version: 31.05.2021)

Project Coordinator:

Alma Mater Studiorum – University of Bologna (Italy)

Project Partners:

University of Köln (Germany)

Catholic University of the Sacred Heart – Milan (Italy)

University of Rouen (France)

University of Salamanca (Spain)

University of Uppsala (Sweden)

Un nuevo "Vocabulario básico" de latín

Guido Milanese[1]

1. Estado de la cuestión

El proyecto de una Certificación Lingüística estándar, tal y como ha sido diseñada por EULALIA, requiere un diccionario de vocabulario básico. La oferta en este campo es amplia (véanse por ejemplo los instrumentos altamente refinados que ofrece el Proyecto Perseus[2] o la lista proporcionada por el Dickinson College (equipo liderado por Christopher Francese)[3]. No obstante, la lista del Dickinson College se limita a 1000 palabras, mientras que el proyectoEULALIA requiere un mayor número de palabras para los niveles superiores. Además, estas listas no tienen información sobre las fuentes utilizadas para componerlas y los miembros del proyecto EULALIA acordaron proporcionar a los potenciales estudiantes una la lista preparada a partir de los textos estándar de latín clásico, esto es, los textos que forman parte del "canon" real adoptado en las escuelas europeas.

Por esta razón, el antiguo y venerable *Vocabulaire de base du latin*, publicado en 1984 por un equipo de Besançon (A.R.E.L.A.B.), que se sirvió del *Dictionnaire fréquentiel* publicado por el grupo LASLA de Liea en 1981, puede seguir siendo utilizado como base fiable con la que construir un nuevo vocabulario básico adecuado a las necesidades de la enseñanza del latín en las escuelas y universidades europeas. Las ventajas de esta obra son:

1. cantidad de palabras: 1600 palabras son compatibles con el proyecto EULALIA.
2. una lista clara de los textos latinos empleados para preparar el vocabulario:

> «Le corpus analysé dans cet ouvrage comprend, en partie ou en totalité, les oevres de Catulle, César, Cicéron, Horace, Juvénal, Ovide, Perse, Properce, Quinte-Curce, Salluste, Sénèque, Tacite, Tibulle, Tite-Live, Virgile, Vitruve»[4].

Ninguna lista es perfecta, claro está, y la ausencia de autores como Terencio o Lucrecio puede ser poco satisfactoria, pero de cara a preparar una lista de palabras "neutra" de latín tardo-republicano y altoimperial la selección de autores es apropiada. Quien suscribe estas páginas considera que el latín medieval puede proporcionar a los maestros una

[1] EULALIA – European Latin Linguistic Assessment.
[2] https://www.perseus.tufts.edu/hopper/help/vocab
[3] http://dcc.dickinson.edu/vocab/core-vocabulary
[4] G. Cauquil - J.Y. Guillaumin, *Vocabulaire de base du latin (alphabétique, fréquentiel, étymologique)*, A.R.E.L.A.B., Besançon 1984, p. 4.

enorme cantidad de obras muy efectivas para la enseñanza del latín, pero también en ese caso el vocabulario básico puede completarse fácilmente.

2. Escaneo y OCR del vocabulario de Besançon

El vocabulario franco-belga fue publicado como libro[5]. Desgraciadamente, la calidad de la impresión original es la de un libro impreso en los años 80 usando un diseño de impresión camera-ready preparado con los dispositivos del momento. La lista de palabras fue escaneada utilizando un escáner estándar HP OfficeJet Pro 7720 y el OCR fue realizado mediante el conocido programa de OCR *tesseract*, con la opción *latin*[6]. La lista fue corregida manualmente, pero para mi sorpresa incluso la primera remesa de resultados de *tesseract* era ya muy buena.

3. Etiquetado morfológico

La lista final proporcionada por el vocabulario de Lieja-Besançon es una mera lista de palabras y frecuencias, con algunas notas añadidas puntualmente. Para utilizarlacomo lista de un auténtico léxico de base, obviamente era necesario añadir el oportuno etiquetado morfológico. Para este propósito, utilicé *treetagger*, el programa morfológico diseñado por Helmut Schmid en el proyecto TC del Instituto de Lingüística Computacional de la Universidad de Stuttgart y actualmente alojado en la Universidad de Múnich, Alemania[7]. Ejecutar *treetagger* en una lista de palabras es un enfoque con una probabilidad de error elevada: el *parser* no puede realizar un análisis contextual y entradas como, por ejemplo, *quam* son evidentemente opacas[8]. El análisis morfológico ofrecido por *treetagger* fue revisado manualmente. Las notas de A.R.E.L.A.B. fueron utilizadas para corregir el etiquetado *treetagger* allí donde divergía de la categoría morfológica ofrecida por la nota A.R.E.L.A.B.

4. El problema de las palabras funcionales

Durante una reunión del equipo de EULALIA, Mélanie Lucciano (Université de Rouen-Normandie) observó que una lista de "pura" frecuencia no era apropiada como

---

[5] La lista de palabras sin nada más ha sido reimpresa en otras obras recientes, como Paolo Lamagna, *Il lessico latino di base*, Bompiani, Milano 1999.

[6] Sobre *tesseract* véase G. Milanese, *Filologia, letteratura, computer. Idee e strumenti per l'informatica umanistica*, Vita e Pensiero, Milano 2020, pp. 103-107, y la documentación desiponible en línea en https://github.com/tesseract-ocr. El programa es FOSS (*Free and Open Source Software*) y fue ejecutado en un sistema operativo Linux Mint 20.1

[7] Véase https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/, y https://www.cis.lmu.de/~schmid/ para bibliografía reciente.

[8] Utilizar la creación más reciente de Schmid, llamada 'RNNTagger' no supone una diferencia sustancial; véase https://www.cis.lmu.de/~schmid/tools/RNNTagger/.

instrumento para la didáctica de una lengua, porque las palabras funcionales ("mots-outils", "morfemas gramaticales", "palabras vacías"…) son muy comunes en las lenguas y se encuentran en el rango más alto de una lista de frecuencia. En consecuencia, si adoptamos una lista de 700 palabras para un nivel *A1*, los estudiantes aprenderían una enorme cantidad de conjunciones y preposiciones en detrimento de verbos, sustantivos, etc.

Las clases morfológicas recogidas por *treetagger* son las siguientes:

- ADJ
- ADJ:NUM
- ADV
- CC
- CS
- DET
- DIMOS
- ENCL
- ESSE
- EXCL
- FW
- INDEF
- INT
- N
- POSS
- PREP
- PRON
- REL
- V

Dado que se trata de enseñanza de la lengua y el motivo de adaptar la lista original de A.R.E.L.A.B. es evitar demasiadas palabras funcionales, yo propondría considerar "palabras funcionales" las siguientes categorías: conjunciones, preposiciones y clíticos.

Un sencillo programa SNOBOL4 ejecutado en la lista A.R.E.L.A.B. produce los siguientes resultados[9]:

- de la lista completa (1635 palabras) 78 son "palabras funcionales", es decir, el 4,1260 %
- de las primeras 600 palabras 55 son "palabras funcionales", es decir, el 9,100%
- de las primeras 400 palabras, 45 son "palabras funcionales", es decir, el 11,100%

Además, el 57% de las "palabras funcionales" se encuentran recogidas entre las primeras 400 palabras y casi el 70% entre las primeras 600 palabras. La consecuencia de ello es que la lista de palabras más frecuentes no es apropiada para evaluar la competencia lingüística de los estudiantes, ya que entre las primeras 400-600 posiciones hay

[9] Aunque es un "lenguaje de nicho", snobol4 tiene un buen mantenimiento, en particular gracias a los esfuerzos de Phil Budne: véase https://www.snobol4.org/. La versión más reciente del lenguaje se lanzó en diciembre de 2020. Sobre este lenguaje, particularmente apropiado para análisis sintáctico, véase Milanese, *Filologia, letteratura, computer* cit., pp. 241-244. Snobol4 funciona en Linux, Windows y OSX.

demasiadas “palabras funcionales”. El propósito de la revisión de la lista es equilibrar las palabras funcionales y las palabras semánticas, subiendo algunas palabras semánticas desde una posición más baja de la lista (= palabras más allá del n.º400 y del n.º 600) con el objeto de obtener una porcentaje más equilibrado de palabras semánticas/palabras funcionales. Dicho de otro modo, las primeras 400 y 600 palabras deberían mostrar un porcentaje de palabras funcionales similar al de la lista completa.

La lista fue revisada desplazando algunas palabras funcionales a posiciones más bajas de la lista, pero tomando la precaución de que las palabras funcionales esenciales (como *quia* o *si*) estuvieran recogidas en las listas para principiantes. El trabajo fue realizado con un proceso asistido por ordenador, es decir, comprobando paso a paso el equilibro entre las palabras.

La lista revisada muestra ahora los siguientes porcentajes:

- de las primeras 600 palabras, 46 son “palabras funcionales”, es decir, el 7,4%
- de las primeras 400 palabras, 30 son “palabras funcionales”, es decir, el 7,2 %.

Aunque estos porcentajes no reflejan exactamente los datos de la lista completa, me parece un compromiso razonable –aunque cuestionable, naturalmente, y abierto a ulteriores mejoras- entre la exactitud y las necesidades prácticas de la didáctica.

5. Resultado final

El resultado final fue producido en forma de lista CSV (Comma Separated Values), susceptible de ser leída por programas estándar del tipo de *Calc* de LibreOffice o *Excel* de MSOffice. He añadido un *header* simple:

N, VOX, MORPHO, NOTAE, FREQ

donde N = número, VOX = la palabra recogida en la lista, MORPHO = la etiqueta morfológica, NOTAE = la nota original del *Vocabulaire* de Lieja-Besançon, y FREQ = la frecuencia de la palabra. Por ejemplo:

16, QUIS, PRON, interr., 4555

La información añadida por el *Vocabulaire* ha sido añadida para especificar el tipo de pronombre al que pertenece *quis*. Cuando la información proporcionada por el *Vocabulaire* es la misma de la etiqueta morfológica, queda tácitamente omitida.

Para facilitar la legibilidad de la lista, el archivo PDF se produjo a partir de la lista CSV utilizando *pandoc*[10].

[10] Véase https://pandoc.org.

6. Listas a partir de la lista

Puesto que el propósito de todo esto es poner a disposición de profesores y estudiantes instrumentos apropiados y fáciles de usar, he reordenado las listas de diferentes formas. El primer paso fue obviamente imprimir las listas de las primeras 400 y las primeras 600 palabras.

6.1. Listas alfabéticas

Todas las listas (es decir, la lista completa y las listas de 400 y de 600 palabras) han sido reordenadas alfabéticamente para que profesores y estudiantes puedan localizar cualquier palabra recorriendo la lista en orden alfabético. La lista alfabética completa ha sido transformada también en archivo PDF.

6.2. Listas morfológicas

Todas las listas (es decir, la lista completa y las listas de 400 y de 600 palabras) han sido reordenadas morfológicamente para que profesores y estudiantes puedan localizar cualquier palabra de una determinada categoría morfológica. La lista morfológica completa fue transformada también en archivo PDF.

7. Un GUI de la lista

El autor ha desarrollado una interfaz gráfica de usuario (GUI) muy simple para facilitar el uso de estas listas y se encuentra a disposición de *beta testers* interesados en este proyecto.

8. Codificación de *snobol4* y *bash scripts*

8.1. El *script* de *snobol4*

El *script* de *snobol4* para el cálculo de porcentajes:

```
    P_type = break(',') len(1) break(',') len(1) (break(','). M_type)
    KL =0
    KE =0
loop Line = input :f(loop_end)
    KL = KL +1
    Line ? P_type
    ((leq(M_type,"CC")), (leq(M_type,"CS")), (leq(M_type,"ENCL")),
    + (leq(M_type,"PREP"))) (KE = KE +1) :(loop)
    loop_end Output = KE'/' KL
    Extreme = KE *100
    Percentage = Extreme / KL
```

```
Modulus = Remdr(Extreme,KL)
Terminal = Percentage'.' Modulus
end
```

La primera línea declara un patrón: en una determinada línea CSV, el cursor se desplaza a la primera coma, luego a un carácter, repite los mismos pasos, y al tercer intento salva la información como tipo morfológico. Luego inicializamos el contador de líneas (KL) y el contador de palabras vacías (KE) en cero[11]. El bucle lee todas las líneas a partir de la salida estándar (stdout); si la clase morfológica es una palabra vacía el contador KE se incrementa en 1. Después del bucle, una serie de sencillas operaciones matemáticas producen el resultado deseado.

El mismo programa se ejecuta en las primeras 400-600 palabras, añadiendo un límite:

```
Lt (KL,601)
```

En este caso el bucle terminará en la línea número 600 (lt = 'less than').

8.2. Los *bash scripts*

Los *bash scripts* utilizados para ordenar las listas en secuencia alfabética o morfológica son simples *sort scripts*, como el siguiente:

```
sort -t, -k3,3 -k2 frequentia-rev-1-400.csv\

> frequentia-rev-1-400-morph.csv
```

Esta instrucción dispone la lista de 400 palabras sobre la base del tercer campo (la clase morfológica) y luego sobre la base del segundo (= alfabéticamente), escribiendo el resultado en un nuevo archivo (frequentia-rev-1-400-morph.csv). Se usa una coma para separar campos ("-t," o "-t','").

8.2. De CSV a PDF usando *pandoc*

Pandoc es un programa de línea de comandos muy eficiente, capaz de transportar de muchos formatos actualmente en uso (por ejemplo, DOCS, LATEX, markdown) a otros formatos[12]. La versión más reciente de pandoc puede leer un archivo CSV y transformarlo en otros formatos, incluido PDF (ejecutando una llamada silenciosa a TEX).

Se trata de esta simple instrucción:

```
pandoc -f csv frequentia-format-rev-morph.csv\
```

[11] Este paso no es necesario pero la inicialización de datos hace que así el programa lea y entienda más fácilmente.
[12] Véase https://pandoc.org/.

```
-s -o frequentia-format-rev-morph.pdf
```

Pandoc lee un archivo CSV (-f significa “from”) y produce un archivo PDF. El mismo esquema se aplica a todos los demás archivos.

Referencias

Cauquil, G. - J.Y. Guillaumin, *Vocabulaire de base du latin (alphabétique, fréquentiel, étymologique)*, A.R.E.L.A.B., Besançon 1984.

Delatte, L., - Et. Evrard, - S. Govaerts, - J. Denooz, *Dictionnaire fréquentiel et index inverse de la langue latine*, L.A.S.L.A. (Laboratoire d’analyse statistique des langues anciennes), Université de Liège, Liège 1981.

Lamagna, Paolo, *Il lessico latino di base*, Bompiani, Milano 1999.

Milanese, Guido, *Filologia, letteratura, computer. Idee e strumenti per l’informatica umanistica*, Vita e Pensiero, Milano 2020.