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

(German 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)

Eine neue Latein-Grundlagenwortkunde

Guido Milanese*

1. Stand des Projekts:

Das Projekt einer Sprachzertifizierung wie EULALIA erfordert ein Verzeichnis des Kernwortschatzes. In diesem Bereich besteht ein breites Angebot, s. z.B. die spezialisierten Hilfsmittel, die vom Perseus-Project[1] angeboten werden, oder die Vokabelliste, die das Team des Dickinson College um Christopher Francese zur Verfügung stellt.[2] Allerdings ist die Liste des Dickinson College auf 1000 Wörter beschränkt, während das EULALIA-Projekt im Hinblick auf höhere Lernniveaus eine größere Wörterzahl erfordert. Zudem geben diese Listen keine Informationen über die Quellen, die bei ihrer Erstellung verwendet wurden. Das EULALIA-Team ist dagegen übereingekommen, den Schülerinnen und Schülern ein Wörterverzeichnis zur Verfügung zu stellen, das auf typischen, ‚klassischen' lateinischen Texten beruht, d.h. auf Texten, die tatsächlich Teil des Kanons in europäischen Schulen sind.

Aus diesem Grund ist das altehrwürdige *Vocabulaire de base du latin*, das 1984 von einem Team aus Besançon (a.r.e.l.a.b.) publiziert wurde, immer noch eine verlässliche Grundlage für die Erstellung eines neuen Kernvokabulars, das auf die Bedürfnisse des Lateinunterrichts an europäischen Schulen und Universitäten abgestimmt ist. Das *Vocabulaire de base du latin* beruht seinerseits auf dem *Dictionnaire fréquentiel,* das 1981 von der Arbeitsgruppe LASLA aus Lüttich erstellt wurde. Es bietet folgende Vorteile:

1. Die Zahl von 1600 Vokabeln stimmt mit den Erfordernissen des Eulalia-Programms überein.
2. Es legt eine klar definierte Liste von Autoren zugrunde:

> «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»[3]

Natürlich ist keine Wörterliste perfekt, und das Fehlen von Autoren wie Terenz und Lukrez kann man zu Recht beklagen. Dennoch ist die Autorenauswahl für eine ‚neutrale' Wörterliste des spätrepublikanischen und frühkaiserzeitlichen Latein angemessen. Der Verfasser glaubt, dass das Mittellatein ebenfalls eine große Zahl von Texten für einen effektiven Lateinunterricht bietet. Auch hier kann das Kernvokabular leicht eingesetzt werden.

2. Scan und OCR der Wortliste aus Besançon

---

* Eulalia: European Latin Linguistic Assessment.

[1] https://www.perseus.tufts.edu/hopper/help/vocab.

[2] http://dcc.dickinson.edu/vocab/core-vocabulary.

[3] 3G. Cauquil and J.Y. Guillaumin, *Vocabulaire de base du latin (alphabétique, fréquentiel, étymologique)*, ARELAB, Besançon 1984, S. 4.

Die französisch-belgische Wortkunde wurde in Buchform publiziert.[4] Die Qualität des Originals entspricht leider dem Standard eines Buchs der 80er Jahre, das mit Hilfe einer camera-ready-Vorlage und den technischen Möglichkeiten der Zeit produziert wurde. Die Wortliste wurde mit einem Standard HP OfficeJet Pro 7720 gescannt, und die OCR Texterkennung mit dem bekannten Programm *tesseract* und der Option „Latein" durchgeführt.[5] Die Wortliste wurde danach manuell korrigiert, aber überraschenderweise war schon das erste mit *tesseract* erzielte Ergebnis sehr gut.

3. Morphologische Markierung

Die endgültige Liste, die sich aus dem Liège–Besançoner Lexikon ergibt, ist eine reine Aufzählung von Wörtern und Angaben zu ihrer Häufigkeit, hier und da ergänzt durch einzelne Anmerkungen. Um diese Liste für ein wirkliches Kernlexikon nutzen zu können, ist es nötig, eine geeignete morphologische Markierung hinzuzufügen. Ich habe dafür *treetagger* genutzt, jenes Programm, das von Helmut Schmid im Rahmen des TC-Projekts am Institut für Computerlinguistik der Universität Stuttgart entwickelt wurde und nun von der LMU München betreut wird.[6] *Treetagger* auf eine Wörterliste anzuwenden, ist relativ fehleranfällig. Der Parser kann keine Kontextanalyse durchführen, und Einträge wie *quam* sind notorisch uneindeutig.[7] Die von *treetagger* durchgeführte Analyse wurde manuell überprüft: Ungefähr 15% der Einträge waren erwartungsgemäß falsch. Zur Korrektur wurden die Einträge von ARELAB verwendet, falls sich die Ergebnisse von *treetagger* von der Kategorie, die der ARELAB-Eintrag aufwies, unterschieden.

4. Das Problem der Funktionswörter

In einem Treffen des EULALIA-Teams merkte Mélanie Lucciano (Université de Rouen-Normandie) an, dass eine nur auf Häufigkeit beruhende Wortliste für den Sprachunterricht wenig geeignet sei, weil sogenannte Funktionswörter in jeder Sprache sehr häufig sind und daher ganz oben in der Liste erscheinen. Wenn wir eine Liste von 400 Wörtern für das Niveau A1 annähmen, würden die Schülerinnen und Schüler daher eine große Zahl von Konjunktionen und Präpositionen lernen müssen, während Substantive, Verben etc. weniger vertreten wären.

Treetagger verwendet folgende morphologische Klassifizierungen:

- ADJ

[4] Die reine Wortliste wurde in neueren Werken nachgedruckt, z.B. in Paolo Lamagna, *Il lessico latino di base,* Bompiani, Milano 1999.

[5] Zu *tesseract* s. Guido Milanese, *Filologia, letteratura, computer. Idee e strumenti per l'informatica umanistica*, Vita e Pensiero, Milano 2020, S. 103-107, und die Online-Dokumentation unter https://github.com/tesseract-ocr. Es handelt sich um ein FOSS-Programm (Free and Open Source Software), das auf einem Linux Mint 20.1-System verwendet wurde.

[6] S. https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ und https://www.cis.lmu.de/~schmid/ for recent bibliography.

[7] Das Programm 'RNNTagger', eine neuere Entwicklung von Schmid, bringt keine wesentlich anderen Ergebnisse, s. https: //www.cis.lmu.de/~schmid/tools/RNNTagger/.

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

Da die ursprüngliche ARELAB-Wörterliste dahingehend modifiziert werden soll, dass eine Häufung von Funktionswörtern zum Nutzen des Sprachunterrichts vermieden wird, schlage ich vor, folgende Wortarten als Funktionswörter zu definieren: Konjunktionen, Präpositionen und Enklitika.

Die Anwendung eines einfachen SNOBOL4-Programms auf die ARELAB-Liste ergibt folgende Ergebnisse[8]:

- Bezogen auf die Gesamtliste (1635 Wörter) ergeben sich 78 Funktionswörter, d.h. 4.1260 %.
- Unter den ersten 600 Wörtern finden sich 55 Funktionswörter, d.h. 9.100 %.
- Unter den ersten 400 Wörtern finden sich 45 Funktionswörter, d.h. 11.100 %.

Es finden sich also unter den ersten 400 Lemmata bereits 57% der Funktionswörter und unter den ersten 600 Wörtern fast 70 %. Das bedeutet, dass eine auf Häufigkeit basierende Liste nicht dazu geeignet ist, die Sprachkompetenz von Schülerinnen und Schülern zu überprüfen, weil sich unter den 400 bzw. 600 häufigsten Wörtern zu viele Funktionswörter befinden. Das Ziel der überarbeiteten Liste ist es, eine Ausgewogenheit zwischen Funktionswörtern und semantischen Bedeutungsträgern zu erreichen. Dafür werden einige semantisch bedeutsame Wörter, die nicht unter die 400 bzw. 600 häufigsten Lemmata fallen, in die Liste

[8] Obwohl SNOBOL4 "Nischensprache" unter den Programmiersprachen ist, wird sie genutzt und überholt, insbesondere dank der Arbeit Phil Budnes, s. https://www.snobol4.org/. Die aktuelle Version der Programmiersprache, die sich besonders für die Textanalyse eignet, wurde im Dezember 2020 zur Verfügung gestellt. Zu dieser Sprache siehe
Milanese, *Filologia, letteratura, Computer* cit., S. 241-244. *SNOBOL4* läuft auf Linux, Windows, und OSX.

aufgenommen. Mit anderen Worten: Die ersten 400 bzw. 600 Vokabeln sollten etwa den gleichen prozentualen Anteil an Funktionswörtern aufweisen wie die Gesamtliste. In der überarbeiteten Liste wurden in der Folge einige Funktionswörter weiter nach unten verschoben, wobei allerdings sichergestellt wurde, dass essenzielle Funktionswörter wie *quia* oder *si* in der Vokabelliste für das Basisniveau erscheinen. Die Überarbeitung wurde mit einem computergestützten Ansatz durchgeführt, bei dem das Verhältnis von Funktionswörtern und semantischen Bedeutungsträgern schrittweise überprüft wurde. In der überarbeiteten Liste ergeben sich folgende Prozentzahlen:

• Von den ersten 600 Wörtern sind 46 Funktionswörter, d.h. 7.4 %.

• Von den ersten 400 Wörtern sind 30 Funktionswörter, d.h. 7.2 %.

Auch wenn dieser prozentuale Anteil nicht genau dem der vollständigen Liste entspricht (~ 4,1 %), erscheint mir das Ergebnis als ein vernünftiger Kompromiss zwischen Präzision und den praktischen Erfordernissen des Unterrichts. Natürlich kann das Ergebnis hinterfragt und weiter verbessert werden.

5. Endergebnis der Gesamtliste

Das Endergebnis wurde als CSV (Comma Separated Values)-Liste erstellt, so dass sie von Standardprogrammen wie *LibreOffice Calc* or *MSOffice Excel* gelesen werden kann. Ich habe eine einfache Kopfzeile hinzugefügt:

N,VOX,MORPHO,NOTAE,FREQ

wobei N = Nummer, VOX = das gelistete Wort, MORPHO = der morphologische Tag (morphological tag), NOTAE = der ursprüngliche Eintrag in der Wortkunde von Liège–Besançon, und FREQ = die Häufigkeit des Wortes ist. Ein beispielhafter Eintrag wäre folgender:

16,QUIS,PRON,interr.,4555

Einträge aus der Wortkunde von Liège–Besançon werden herangezogen, um die Art des Pronomens zu spezifizieren, zu dem die Form *quis* gehört. Wenn die Information aus der Wortkunde von Liège–Besançon dem morphologischen Tag entspricht, wird sie weggelassen und die NOTAE-Zelle bleibt unausgefüllt. Um die Liste lesbarer zu machen, habe ich aus der CSV-Liste mit Hilfe von *pandoc* eine PDF-Datei erstellt.[9]

6. Weitere aus der Gesamtliste generierte Listen

Um Lehrenden und Studierenden ein leicht einsetzbares Hilfsmittel zur Verfügung zu stellen, habe ich die Wortliste überdies nach verschiedenen Gesichtspunkten angeordnet. Der erste und selbstverständliche Schritt ist ein Ausdruck der ersten 400 bzw. 600 Wörter.

6.1 Alphabetische Listen

[9] Siehe https://pandoc.org.

Alle Listen (d.h. die Gesamtliste und die Liste der ersten 400 bzw. 600 Wörter) wurden alphabetisch sortiert, so dass Lehrende und Lernende ein Wort leicht auffinden können. Es wurde zudem ein PDF-Dokument der Gesamtliste erstellt.

6.2 Morphologische Listen

Alle Listen (d.h. die Gesamtliste und die Liste der ersten 400 bzw. 600 Wörter) wurden morphologisch sortiert, so dass Lehrende wie Lernende alle Wörter einer bestimmten morphologischen Kategorie leicht auffinden können. Es wurde zudem ein PDF-Dokument der morphologisch sortierten Gesamtliste erstellt.

7. Ein graphisches Interface für die Liste

Um die Nutzung der Listen zu erleichtern, hat der Autor ein sehr einfaches graphisches Interface entwickelt. Personen, die an diesem Projekt Interesse haben, können die Beta-Version testen.

8. Programmierungscode: SNOBOL4 und *bash scripts*

8.1 Das SNOBOL4 script

Das snobol4 script für die Errechnung der prozentualen Anteile lautet:

```
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
```

Die erste Kommandozeile gibt ein Muster an: In einer bestehenden CSV-Zeile (z.B. 16,QUIS,PRON,interr.,4555) bewegt sich der Cursor zum ersten Komma, geht dann ein Zeichen weiter und wiederholt dieselben Schritte. Beim dritten Durchlauf speichert das Programm die Information als „morphologischer Typ“. Dann wird die Zeilenzählung (KL) und die Leerwortzählung (KE) auf Null voreingestellt.[10] Der Loop liest alle Zeilen des Standard-Outputs aus. Wenn die morphologische Kategorie ein Leerwort ist, erhöht sich der Leerwortzähler um 1. Nach dem Loop gibt das Programm mittels einiger einfacher mathematischer Operationen das gewünschte Ergebnis aus. Dasselbe Programm wird auf die Listen der ersten 400 bzw. 600 Wörter angewandt, dann aber mit einem zusätzlichen Limit:

lt(KL,601)

In diesem Fall endet der Loop in der 600. Zeile (lt = less than).

8.2 Die bash scripts

Die bash scripts, die für die alphabetische und die morphologische Sortierung der Listen verwendet wurden, sind einfache Sortierungsscripts wie das folgende:

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

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

Dieser Befehl sortiert die Liste der 400 Wörter nach dem 3. Feld (= die morphologische Kategorie) und dann nach dem 2. Feld (=alphabetisch) Das Ergebnis wird dabei in eine neue Datei geschrieben (frequentia-rev-1-400-morph.csv). Ein Komma dient dazu, die Felder voneinander zu trennen. (“-t,” or “ -t',' ”).

8.3 Mit Hilfe von *pandoc* von der CSV-Datei zur PDF-Datei

Pandoc ist ein sehr effektives Befehlszeilenprogramm und ein Dateikonverter, der viele gängige Formate (z.B. DOCX, LATEX, *markdown*) in andere Formate umwandeln kann.[11] Mit der neusten Version von *pandoc* kann also die CSV-Datei gelesen und in eine PDF umgeschrieben werden. (Es ruft im Hintergrund TEX auf). Dies ist der Befehl:

pandoc -f csv frequentia-format-rev-morph.csv\
-s -o frequentia-format-rev-morph.PDFPDF

Pandoc liest eine CSV-Datei (-f steht für ″from″) and erstellt eine PDF-Datei. Dasselbe Prinzip wird bei anderen Formaten angewendet.

Bibliographie

Cauquil, G. and J.Y. Guillaumin, Vocabulaire de base du latin (alphabétique, fréquentiel, étymologique), ARELAB, Besançon 1984.

[10] Dieser Schritt ist nicht notwendig, aber die Dateninitialisierung macht das Programm verständlicher und leichter lesbar.
[11] S. https://pandoc.org/.

Delatte, L., Et. Evrard, S. Govaerts, and 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.