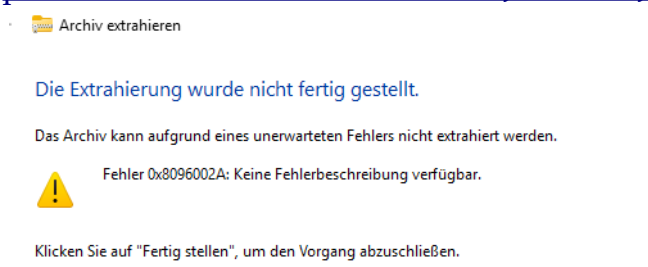
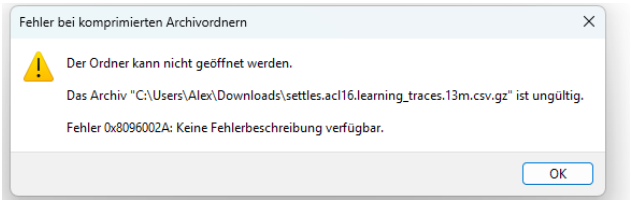
Projekt Vokabellernen

Ich wollte ein Thema rund um Fremdsprachenerwerb.
Leider gab es da nicht viele Daten und die bekannten Studien zB. zur Vergessenskurve nach Ebbinghaus bieten keine öffentlichen Daten, oder die Experimente sind nur mit einer handvoll Lernern durchgeführt worden.

Ich hätte gerne es anhand von meinen FinnischLernern gemacht (auch nicht viele, aber dafür länger als nur 2 Wochen), aber das geht zeitlich nicht und die Programmierung ist nicht gefragt, sondern die Datenaufbereitung.

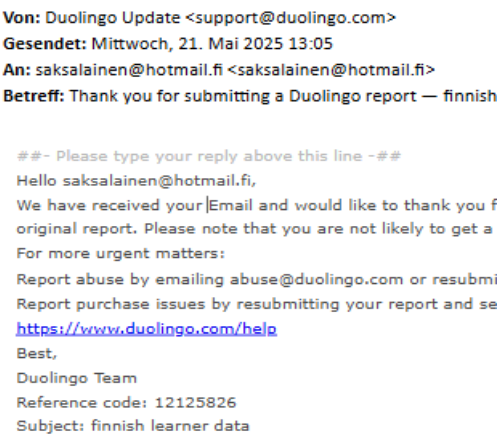
Daher suchte ich und fand in Harvard Zahlen von Duolingo, leider ist der Download irgendwie kaputt und das ZIP lässt sich nicht öffnen (zur Kontrolle, ob das Problem vorm Bildschirm sitzt, bat ich Alexander die Daten zu holen, aber auch er erhielt exakt den gleichen Fehler)
https://dataverse.harvard.edu/file.xhtml?persistentId=doi:10.7910/DVN/N8XJME/UEPJVH&version=1.0





Daten sind selten einmalig, die Datei hieß "learning-traces.13m.csv" also suchte nach dem Dateinamen. Und er war doch in Kaggle, wo ich mit meinen Keywords nichts gefunden hatte(ich hatte immer vocabulary dabei).
https://www.kaggle.com/datasets/aravinii/duolingo-spaced-repetition-data/data?select=learning_traces.13m.csv

Weiterhin habe ich DuoLingo angeschrieben, ob ich Lernerdaten für Finnisch (Zielsprache) erhalten könnte, aber ich rechne nicht ernsthaft mit einer positiven Antwort und schon gar nicht zeitnah.


Von: Duolingo Update <support@duolingo.com>
Gesendet: Mittwoch, 21. Mai 2025 13:05
An: saksalainen@hotmail.fi <saksalainen@hotmail.fi>
Betreff: Thank you for submitting a Duolingo report — finnish

##- Please type your reply above this line -##
Hello saksalainen@hotmail.fi,
We have received your Email and would like to thank you f
original report. Please note that you are not likely to get a
For more urgent matters:
Report abuse by emailing abuse@duolingo.com or resubmi
Report purchase issues by resubmitting your report and se
https://www.duolingo.com/help
Best,
Duolingo Team
Reference code: 12125826
Subject: finnish learner data

Also arbeite ich mit den vorhandenen und habe mir die Daten schon einmal grob angeschaut.

Ich habe in der großen Datenmenge einige Infos über die Lerner
- welche Sprache lernen sie
- was ist ihre Muttersprache
- wie oft haben sie ein Wort diesmal repräsentiert bekommen (und richtig gehabt)
- wie oft haben sie das Wort in all ihren Lernsitzungen schon gesehen und gekonnt
- wie lange ist es her, dass die letzte Wiederholung war

Über das Wort weiß ich auch, welche Sprache es ist und welche Wortart es ist. Theoretisch steht auch in den Daten, in welcher gebeugten Form es geübt wurde.
Beispiel:

"los/el<det><def><m><pl>"

Die bedeutet: gelernt wurde *los*, das kommt von *el,* ist ein bestimmter(*def)* Artikel (*det)* für maskulin(*m*) Plural(*pl*).
Ich werde in der weiteren analyse überlegen, ob ich diese Informationen nutze oder nur die Grundform verwende. Denn es wurde hauptsächlich Englisch von nicht-Englisch-Muttersprachlern gelernt bzw. die EnglischMuttersprachler lernten hauptsächlich Spanisch. Aber der Formenreichtum unterscheidet sich gewaltig, was sicher für EnglischSprecher ein Lernhindernis ist. Was ich versuchen werde anhand der fehlerhaften Wiederholungen darzulegen. (**These1**)

Weiterhin ist davon auszugehen, dass es sich bei ENG-SPA und *-ENG um das Erlernen einer ersten Fremdsprache geht. Während es sich bei ENG-DEU, ENG-ITA oder ENG-FRA eher um eine zweite oder dritte Fremdsprache handelt, hier würde ich gerne sehen, ob es einfacher wird. (**These2**)

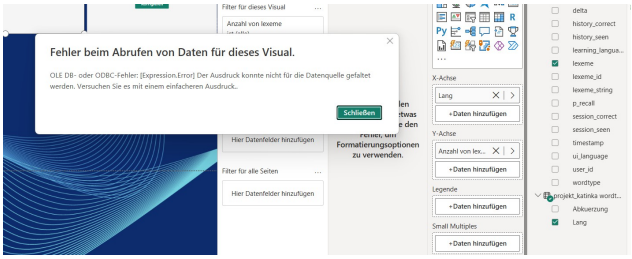
Leider habe ich keine nicht-englisch-Muttersprachler, die etwas anderes als Englisch lernen. Dies wäre sowohl für These1 (zB. Durch einen Vergleich DEU-SPA und ENG-SPA sehr interessant, da sie lexikalisch etwa gleich entfernt zum Spanischen sind, aber grammatikalisch liegt das Deutsche deutlich näher) oder auch für These2 hilfreich.

Weiterhin möchte ich überprüfen, ob sich Nomen(sehr konkret und „greifbar") leichter lernen lassen, als Verben oder gar Präpositionen, die oft nur überschneidende Übersetzungen bieten (on the table=auf dem Tisch, aber on the street= in der Straße; the building by the busstop= das Gebäude bei der Bushaltestelle, aber by chance-zufällig, by the river-am Fluß usw usw) (**These3**).

ICH BITTE UM RÜCKMELDUNG, ob dies den Ansprüchen genügt.

Technisch gesehen möchte ich mit direktQuery in PowerBI arbeiten, weil mit über 12 Millionen Datensätzen das Programm auf meinem Rechner nicht läuft (ich hatte versucht die Datei zu bearbeiten in PowerBI, das ist abgebrochen).
Streamlit macht zu oft nicht so, wie ich will, daher möchte ich es zur Darstellung vermeiden. In einem JupyterNotebook mit Pandas(vielleicht Polars) und Plotly könnte ich es mir auch vorstellen.

Da PowerBI nur bedingt komplexere SQLAbfragen selbst gestaltet, würde ich die Daten in Views fertig zusammenstellen und auch Trigger für Functions erstellen, dass die Views kontinuierlich upgedatet werden, wenn neue Daten hereinkämen(hier natürlich nicht, aber tun wir mal so). Außerdem kann ich mehr als „nur“ PowerBI nachweisen.



Gebraucht wäre:
select count(distinct(v.lexeme), w.lang from vocabel_learning v join wordtype w on v.wordtype=w.abkuerzung group by w.lang

IST DAS GUT SO? VERBESSERUNGSVORSCHLÄGE?

Stand: 22.05.25