**Ergebnis Levensthein für Phontype Vergleich**

Stand 12.7.2023

**Data**

Bisher haben wir 16000 Einzelvergleiche durchgeführt, also 8 Items mit je ca 2000 Orten = 16000.   
Bei 99.9939% der Fälle („ClearLex“) gibt es in einem der Datensätze nur einen Lextype (l) (Gegenüber ein oder mehreren Lextyps im anderen) zB lA und lA,lB. Bei 93 Fällen („MultiLex) gibt es mehr als einen Lextyp in beiden Datensätzen.

Für die „CleanLex“ Fälle verwende ich eine „BestFIT“ Methode: Für den Ort wird die Levensthein Distanz für alle Päärchen der Phontyps (p) berechnet (pA mit pA, pA mit pB) und das beste Ergebnis wird für den Ort genommen. Die Annahme hierbei ist, dass der Phontyp mit der höchsten Ähnlichkeit jener ist, welcher zu dem übereinstimmenden Lextyp gehört.

|  |  |  |  |
| --- | --- | --- | --- |
| dwa\_lextype | mau\_lextype | dwa\_phontype | mau\_phontype |
| Ameise | Ameise, Bärams(e)le | amese | amese, bäramsle |
| Bärams(e)le | Bäramese | bäramsel | Bäramese |
| Gugummer, Gurke | Gugummer | gagummere, gurke | Gugummere |

Tabelle 1: „CleanLex“ Fälle mit „BestFIT“ Methode

Alternativ habe ich eine Funktion geschriebene, welche die Phonotys an der Position des übereinstimmenden Lextyps vergleich (byPOS).

|  |  |  |  |
| --- | --- | --- | --- |
| dwa\_lextype | mau\_lextype | dwa\_phontype | mau\_phontype |
| Ameise[1] | Ameise[1], Bärams(e)le | amese[1] | amese[1], bäramsle |
| Bärams(e)le[1] | Bäramese[1] | bäramsel[1] | Bäramese[1] |
| Gugummer[1], Gurke | Gugummer[1] | gagummere[1], gurke | Gugummere[1] |

Tabelle 2: „CleanLex“ Fälle mit „byPOS“ Methode

**Vergleich der Methoden**

Bei einem Vergleich der Ergebnisse mit beiden Methoden für Ameise (für die „ClearLex“) gibt es Abweichungen

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | BestFIT | | byPOS | |  |
| lex\_name | n\_entries | reliability | n\_entries | reliability | Difference |
| Ameise | 1661 | 0.96284725 | 1661 | 0.95144706 | 0.01140019 |
| Ameisele | 61 | 0.92811606 | 61 | 0.90947827 | 0.01863779 |
| Bärams(e)le | 63 | 0.91635802 | 63 | 0.90665785 | 0.00970017 |
| Wurmeisle | 56 | 0.96602183 | 56 | 0.92286706 | 0.04315477 |
| Seichemes | 38 | 0.89119484 | 38 | 0.89119484 |  |
| Pfetzemes | 17 | 0.97112299 | 17 | 0.96458705 | 0.00653594 |
| Wurmeise | 1 | 0.875 | 1 | 0.875 |  |
| Bäramese | 2 | 1 | 2 | 1 |  |
| Bäramm(e)le | 2 | 1 | 2 | 1 |  |
| Oremeise | 2 | 1 | 2 | 1 |  |
| Klamoasa | 2 | 0.875 | 2 | 0.875 |  |
| Rotemse | 2 | 1 | 2 | 1 |  |
| Brunzenesse | 0 | NaN | 0 | NaN |  |
| Gailimes | 1 | 1 | 1 | 1 |  |
| Seichemsle | 1 | 1 | 1 | 1 |  |
| Wurmbasle | 1 | 1 | 1 | 1 |  |

Tabelle 3: Vergleich der Ergebnisse von „BestFIT“ und „byPOS“

**Beispiele für unterschiedliche Ergebnisse der Methoden**

|  |  |  |  |
| --- | --- | --- | --- |
| dwa\_lextype | mau\_lextype | dwa\_phontype | mau\_phontype |
| Pfetzemes | Pfetzemes | petsemese | petsemede, petsemese |
| Wurmeisle | Wurmeisle | wurmeisl, wurmeisle | wurmeisle |
| Ameise, Bärams(e)le, Wurmeisle | Wurmeisle | ameise, baramsle, bäramsle, wurmeisle | Wurmeisle |

Tabelle 4: Beispiele für den Phontyp Vergleich mit „BestFIT“

|  |  |  |  |
| --- | --- | --- | --- |
| dwa\_lextype | mau\_lextype | dwa\_phontype | mau\_phontype |
| Pfetzemes[1] | Pfetzemes[1] | petsemese[1] | petsemede[1], petsemese[2] |
| Wurmeisle[1] | Wurmeisle[1] | wurmeisl[1], wurmeisle[2] | wurmeisle[1] |
| Ameise, Bärams(e)le, Wurmeisle[3] | Wurmeisle[1] | ameise, baramsle, bäramsle[3], wurmeisle | Wurmeisle[1] |

Tabelle 5: Beispiele für den Phontyp Vergleich mit „byPOS“

Die „BestFIt“ Methode erscheint eine höhere Genauigkeit zu besitzen und ich stabiler gegenüber Fehler bei der Reihenfolge. Allerdings falls ein pB ähnlicher zu pA ist als pA zu pA würde diese Methode die Ergebnisse verfälschen.

Die „byPOS“ Methode erzeugt bei den Beispielen Fehler, ist aber zuverlässiger falls ein pB ähnlicher zu pA ist als pA zu pA.

**Umgang mit „MultiLex“**

In den bisherigen Items gibt es insgesamt aber 97 Fälle (0.0060%), wo „MultiLex“ auftritt. Allerdings sind bei 36 der 97 Fälle sowohl die Lextyps als auch die Anzahl an Lex- und Phontyps identisch. Hier würde die „BestFIT“ Methode zu Fehler führen, da für beide Lextyps jeweils die beste Übereinstimmung aller Phontyp Pärchen gewählt werden würde (siehe Tabelle 6 Geld markiert). Allerdings würde für diese Fälle die „byPOS“ Methode funktionieren.

|  |  |  |  |
| --- | --- | --- | --- |
| dwa\_lextype | mau\_lextype | dwa\_phontype | mau\_phontype |
| Beerdigung, Li(e)cht | Beerdigung, Li(e)cht | beerdigung, liX | beerdigung, leiXe |
| Arschkrätzer, Butte | Arschkrätzer, Butte | arSkratzer, butte | arSkrätzer, butt |
| Geiß, Zick | Geiß, Zick | geis, zig | geis, zeg |

Tabelle 6: „BestFIT“ funktioniert nicht bei mehreren identischen Lextyps

**Ausblick / toDO**

Kann ein pB ähnlicher zu pA sein als pA zu pA. Wenn nicht lässt sich „BestFIT“ sicher anwenden.

Sofern in den übrigen Fällen eine der beiden Methoden greift und diese Fälle klar abgrenzbar sind können wir alles automatisch machen. Wenn nicht müssen wenige Einzelfälle händisch ergänzt werden.