

Synchronic grapheme and phoneme correspondences in German-Dutch cognates

Lingtalk – 22.03.24

Content

1. Scientific background
2. Research questions
3. Cognates
4. Levenshtein Distance
5. Conditional Entropy
6. Data
7. Results
8. Conclusion
9. Current problems & possible solutions

Scientific background(1) - Intercomprehension

Intercomprehension:

- Passive competence (in this case reading) in a non-acquired language that is related to an acquired one
 - German – Dutch
 - Norwegian – Swedish
 - Spanish - Portuguese
- Intercomprehension studies focus on different aspects:
 - similarity of cognates
 - intelligibility
 - Extra-linguistic factors

Scientific background(2) - Similarity and intelligibility

- Cognate similarity:
 - Cognates with a higher degree of similarity are more easily identifiable (Van Bezooien & Gooskens 2007)
 - Measuring the phonological, orthographic, etc. distance between the cognates (Möller & Zeevaert 2015, Gooskens & van Heuven 2020)
- Other indicators: Conditional Entropy (Moberg et al. 2007) and Adaptation Surprisal (Jágrová et al. 2019)
- Intelligibility:
 - Generally test subjects without any knowledge of systemic correspondences (Berthele 2008, Gooskens & van Heuven 2020)

Research questions(1) - In general

- Combining two less studied aspects of intercomprehension:
 - Quantitative synchronic relevance of historically attested correspondences
 - How much knowledge of systemic correspondences is absolutely necessary for a fairly high degree intercomprehension?
- Determining the synchronically present correspondences allows selection of relevant correspondences for
 - automated cognate recognition
 - a set of limited rules necessary to being able to identify a maximal number of cognates
- Intelligibility tests with subjects having been taught this limited set of rules

Research questions(2) - This presentation

- German-Dutch list of cognate verbs
 - Analysis of orthographic differences between German and Dutch cognates
 - Analysis of vowel/diphthong correspondences
 - Analysis of possible asymmetry in correspondences

Cognates

Cognates

- Focus on German and Dutch
 - Two West-Germanic languages
 - High German consonant shift
- Different definitions for the term "cognates":
 - Words that have a common ancestor, i.e. they are defined as historically related words
 - "Translation equivalents that are identical or similar in their orthography across languages" (Dijkstra et al. 2010: p.284)
 - Or a combination of the two (see Heeringa et al. 2014)
 - GE. *Schubblade* - DU. *la*

Levenshtein Distance

Levenshtein Distance(1)

- Regularly used in linguistic research (Heeringa et al. 2014, Möller & Zeevaert 2015)
- Basic Levenshtein distance: indicates the number of operations necessary to transform one word, sentence, etc. into another one
- 3 operation types for basic Levenshtein distance: insertion, deletion, replacement
- GE. *laufen* - DU. *lopen* (EN. *to run*)

GE	L	A	U	F	E	N
DU	L	O	P	-	E	N
COST		1	1	1		

Levenshtein Distance(2) - Unweighted

- No nuances in basic LD - same LD for
 - GE. *laufen* - NL. *lopen* (EN. *to run*)
 - GE. *laufen* - NL. *loeren* (*loeren*: EN. *to lie in wait*)

GE	L	A	U	F	E	N
DU	L	O	E	R	E	N
COST		1	1	1		

Levenshtein Distance (3) - Weighted

- Weighted LD: different operations are accorded specific costs determined by likelihood of operation, leniency towards specific replacements, etc.
- By using a lower weight for the historically attested correspondences <f> - <p> and <au> - <o>, we get a lower overall distance between the two words and they are aligned correctly

GE	L	AU	F	E	N
DU	L	O	P	E	N
COST		0.5	0.5		

Conditional Entropy

Conditional Entropy(1)

- For a given grapheme in German or Dutch, there may be several different corresponding graphemes in the other language, e.g.
 - GE. *Tier* - DU. *dier* (EN. *animal*)
 - GE. *Spiel* - DU. *spel* (EN. *game*)
- The number and frequency of those different corresponding graphemes influence the intelligibility between cognate pairs
- Not necessarily symmetric
 - DU. *blijven* – GE. *bleiben*
 - GE. *Bein* - DU. **bijn* >< DU. *been*

Conditional Entropy(2)

- Conditional entropy measures "the uncertainty, and therefore difficulty of predicting a unit in the native language given a corresponding unit in the non-native language" (Moberg et al. 2007).
- It takes into account three different factors:
 - The probability of a unit appearing in language A
 - The probability of the corresponding unit appearing in language B
 - The probability of those two units aligning
- The higher the entropy, the higher the uncertainty

Conditional Entropy(3) - Example

$$H(X|Y) = - \sum_{x \in X, y \in Y} p(x, y) \log_2 p(x|y)$$

- German diphthong <au> -> two possible correspondences <o> and <ui> in Dutch
- Entropy(DU when GE is given)
= $-(1/8 * 1/4 * \log_2(1/2) + 1/8 * 1/4 * \log_2(1/2))$
= $-(0.125 * 0.25 * -0.301 + 0.125 * 0.25 * -0.301)$
= 0.018
- Entropy(GE when DU is given)
= 0

GE	L	AU	F	E	N
DU	L	O	P	E	N
		1/2			
GE	M	AU	S		
DU	M	UI	S		
		1/2			

Data

Data(1) - Word list

- *Morphy* (Lezius 1996)
- List of 8367 German verbs tagged as infinitives
- Automated translation with Google and Bing (Python *deep-translator* and *translation* packages respectively)
 - Increases probability of finding a (correct) translation
 - Comparison between different translations allows for preselection in case of polysemy
- 6450 successfully translated words, i.e. German verbs that were translated as Dutch verbs

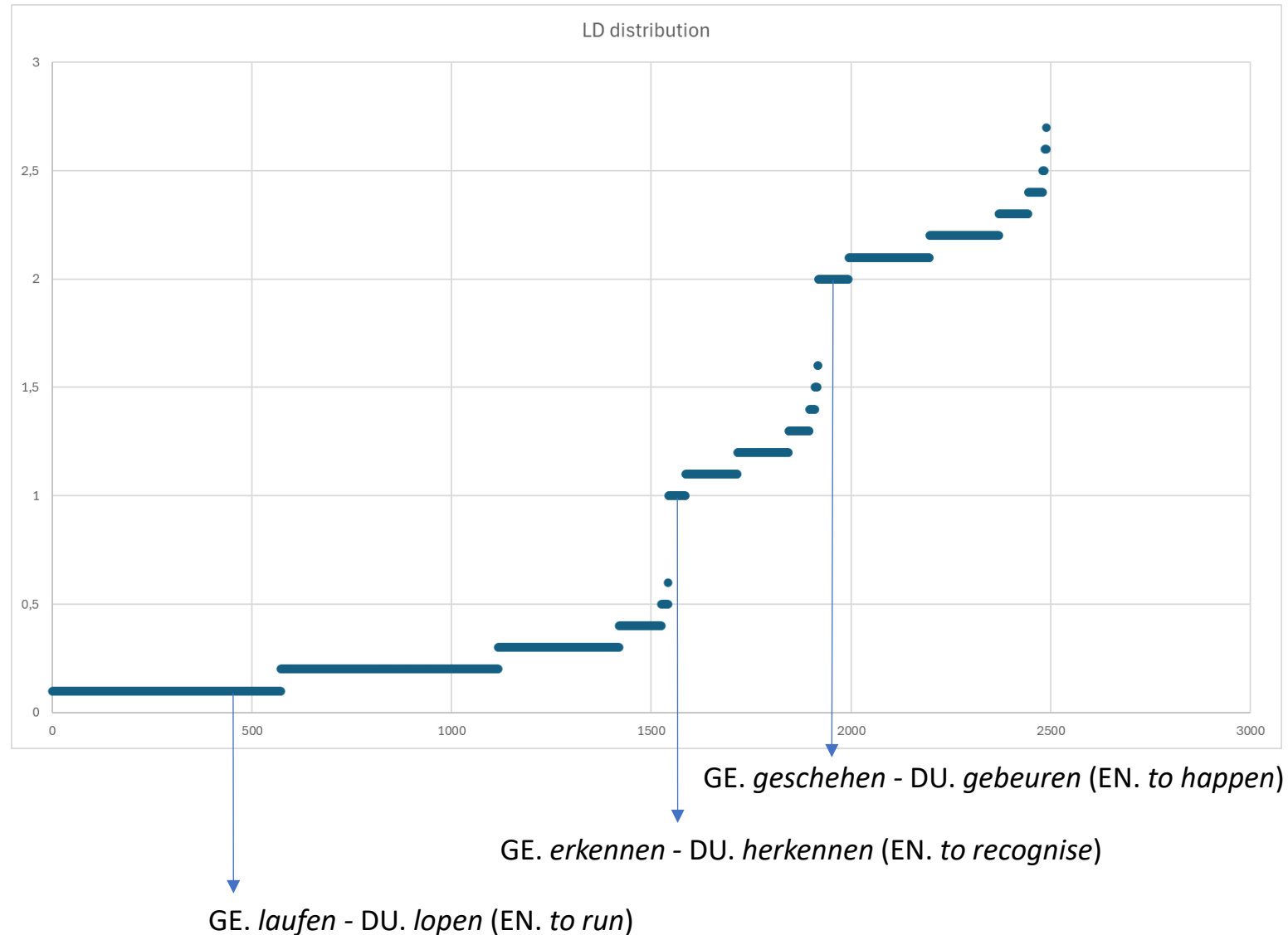
Data(2) - LD weights

- Used weights for word selection:
 - 0: for identical graphemes
 - 0.1 : for historically attested correspondences of consonants, orthographic differences and replacement of a vowel/diphthong with another vowel/diphthong – specific vowel/diphthong correspondences were not taken into account
 - 1: for all other replacements, insertions or deletions
- They are based on orthographic similarity and historically attested correspondences combined from different sources (Rijpma & Schuringa 1969, van Loey 1970, Kuen 1993, Lutjeharms 2014, Lutjeharms & Möller 2014)

Results

Results(1) - LD

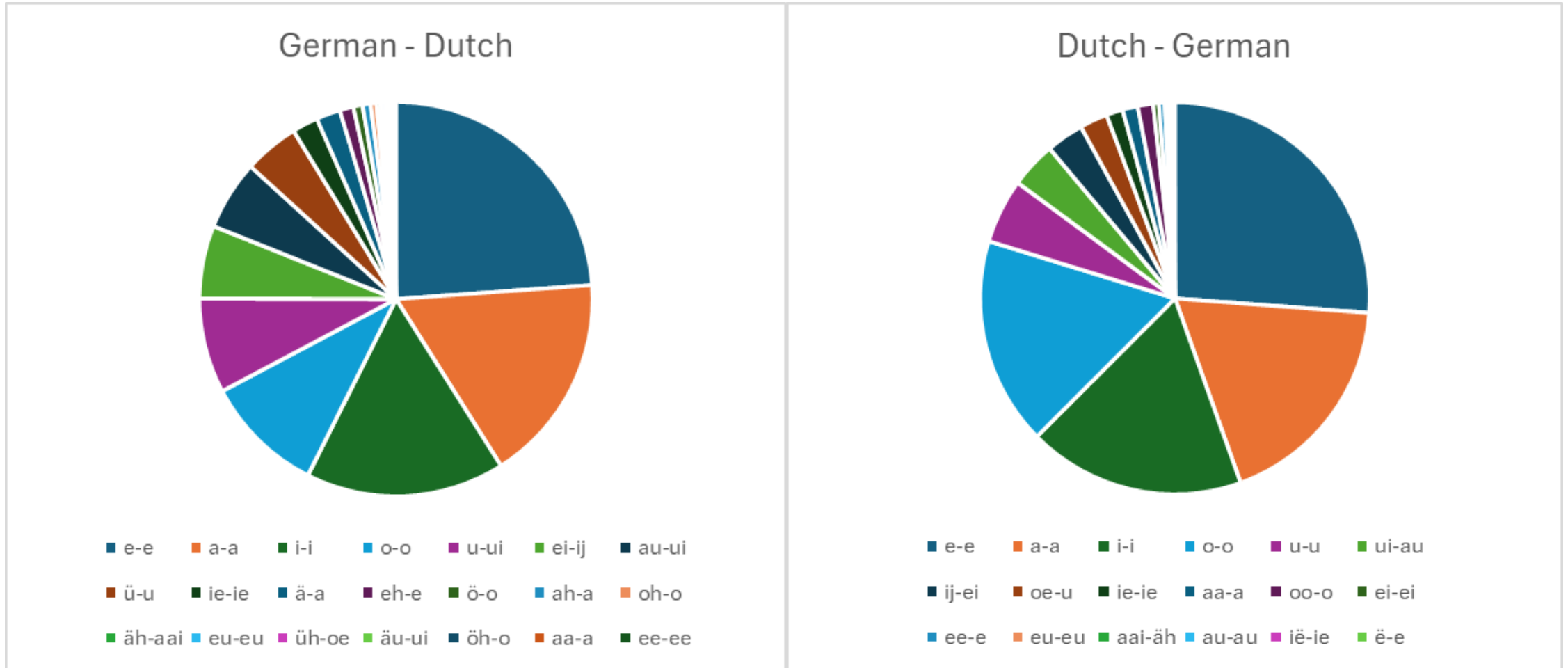
- Lower third of the curve: word pairs containing only historically attested correspondences
- Mostly distances of between 0.1 and 0.3, i.e. there are rarely more than three operations needed to transform one cognate into the other
- Distances between 1.0 and 1.6 are mostly due to prefix differences between the two cognates



Results(2) - Historically attested correspondences

- For the correspondences in both direction (GE – DU and DU – GE), of the 20 most frequent correspondences, only 3 are not the same grapheme in both languages:
 - GE. - DU. <v> (17th)
 - GE. <f> - DU. <v> (19th)
 - GE. <t> - DU. <d> (20th)
- - <v>: consonant change West Germanic -> Old High German
- <f> - <v>: consonant change West Germanic -> Old High German
- <t> - <d>: High German consonant shift
- Typical examples of correspondences caused by the High German consonant shift occur less frequently:
 - GE. <z> - DU. <t> (32nd)
 - GE. <pf> - DU. <p> (64th)

Results(3) - Most frequent correspondences



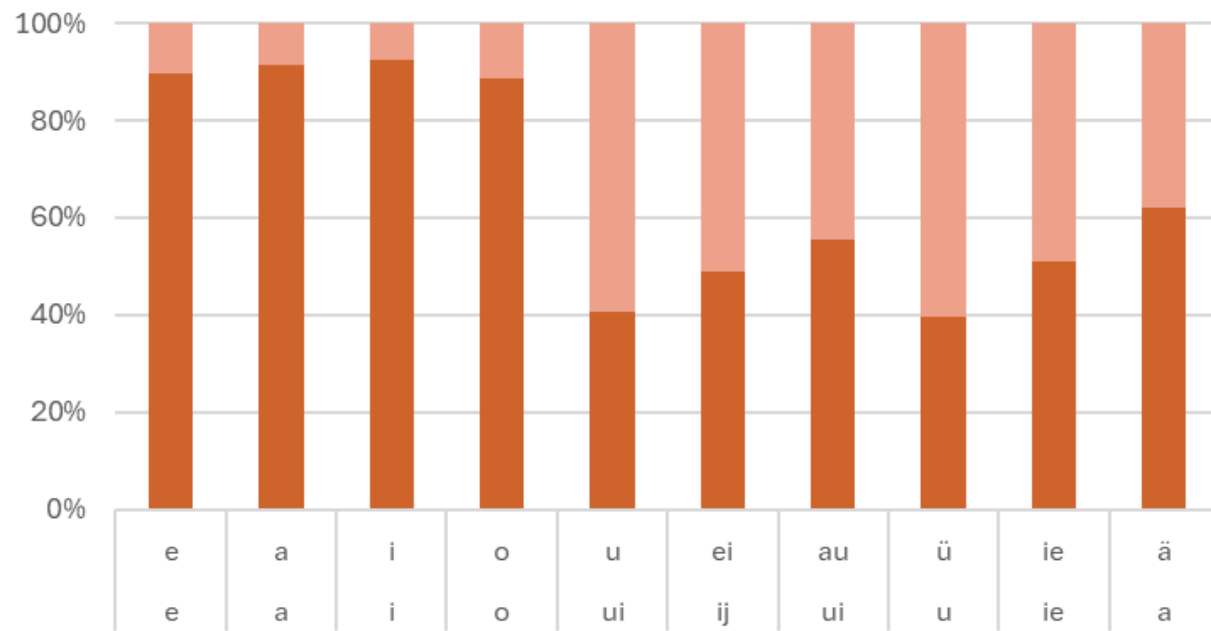
Most frequent correspondences for each of the vowels/diphthongs analysed and their relative frequencies to each other

Results(4) - Uniqueness of correspondences

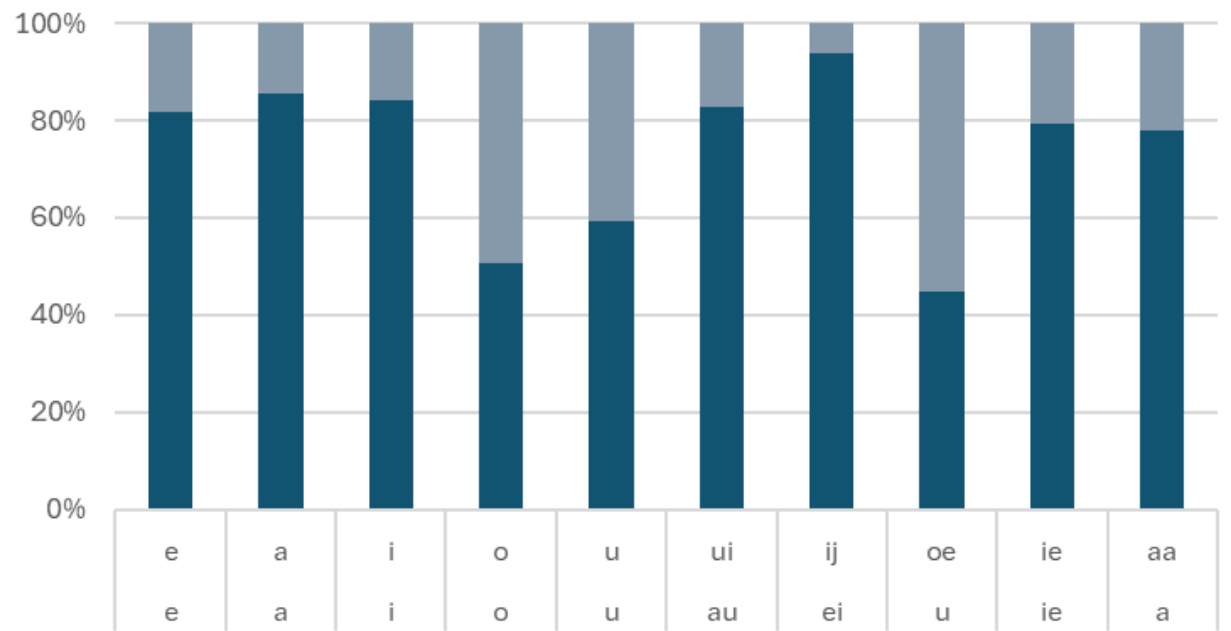
- GE – DU:
 - Average of 4.6
- DU – GE:
 - Average of 5.5
- For only the most frequent correspondences discussed above, these averages go up to
 - 6.5 for GE – DU
 - 10.2 for DU – GE

Results(5) - Most common correspondence

GE - DU



DU - GE



Results(6) - Conditional entropy

- Conditional entropy of vowel/diphthong correspondences
 - GE-DU: **0.797**
 - DU-GE: **0.876**
- Slight asymmetry: vowel/diphthong graphemes are a little bit easier for a Dutch speaker reading German to decipher than the other way around
- But, overall fairly low difference in entropy for both languages. Compare this with the results of Moberg et al. (2007) for Norwegian and Swedish cognates:
 - Swedish-Norwegian: 1.6
 - Norwegian-Swedish: 1.8

Conclusion(1)

- Most differences between the cognates are due to 1, 2 or 3 non-identical graphemes
 - In-word context as an important factor for intelligibility
 - GE. *antworten* – DU. *antwoorden* (EN. *to answer*)
 - GE. *ausschwefeln* – DU. *uitzwavelen* (EN. *to sulphurise*)
 - How often do frequent correspondences cooccur?
 - How many orthographic neighbours exist?
- Most of the correspondences are not based on historically attested correspondences
 - BUT: the presence of one or more of those correspondences can make the mapping of one cognate onto another more difficult
 - Closer analysis of those correspondences needed (e.g. place of articulation – see Möller & Zeevaert 2015)

Conclusion(2)

- A non-negligible amount of asymmetry for a language pair that is often considered a prime candidate for intercomprehension
 - Must be accounted for when creating set of correspondences to be taught (different lists for GE – DU and DU – GE)
 - Due to mergers in German?
 - Asymmetry in consonants?

Current problems and possible solutions

- Prefixes have an influence on the similarity measures and the frequency of correspondences -> repeating stems and repeating prefixes
=> token frequencies
 - Determining prefixes possible by looking for vowels and diphthongs
 - BUT: Where does the prefix end and the stem start?
- Accounting for cooccurring phonemes/graphemes
 - GE. <sch> - DU. <z> if they precede a <w>, e.g.
GE. *schwimmen* – DU. *zwemmen*

Sources

- Berthele, Raphael (2008): Dialekt-Standard Situationen als embryonale Mehrsprachigkeit. Erkenntnisse zum interlingualen Potenzial des Provinzlerdaseins. In: *Sociolinguistica* 22, 87-107.
- Dijkstra, T. et al. (2010): How cross-language similarity and task demands affect cognate recognition. In: *Journal of Memory and Language* 62, 284-301.
- Gooskens, Charlotte/van Heuven, Vincent J. (2020): How well can intelligibility of closely related languages in Europe be predicted by linguistic and non-linguistic variables? In: *Linguistic Approaches to Bilingualism* 10(3), 351-379.
- Heeringa, W. et al. (2014): Modeling Intelligibility of Written Germanic Languages: Do We Need to Distinguish Between Orthographic Stem and Affix Variation? In: *Journal of Germanic Linguistics* 26(4), 361-394.
- Jágrová, Klára et al. (2019): Language models, surprisal and fantasy in Slavic intercomprehension. In: *Computer Speech & Language* 53, 242-275.
- Kuen, E. (1993): Deutsch-Niederländisches Lernwörterbuch.
- Lezius, W. (1996): *Morphy*.
- Lutjeharms, M. (2014): Kapitel 11: Miniporträt Niederländisch. In: Hufeisen, B. / Marx, N. (ed.): Die sieben Siebe. Germanische Sprachen lesen lernen, 2nd edition.
- Lutjeharms, M./Möller, R. (2014): Kapitel 3: Sieb Lautentsprechungen. In: Hufeisen, B. / Marx, N. (ed.): Die sieben Siebe. Germanische Sprachen lesen lernen, 2nd edition.
- Moberg, J. et al. (2007): Conditional Entropy Measures Intelligibility among Related Languages. In: *Proceedings of the 17th Meeting of Computational Linguistics in the Netherlands*, 51-66.
- Möller, R./Zeevaert, L. (2015): Investigating word recognition in intercomprehension: Methods and findings. In: *Linguistics* 53(2), 313-352.
- Rijpma, E./Schuringa, F.G. (1969): *Nederlandse Spraakkunst*.
- Van Bezooien, Renée/Gooskens, Charlotte (2007): Interlingual text comprehension: linguistic and extralinguistic determinants. In: ten Thije, Jan D./Zeevaert, L (Hg.): *Receptive Multilingualism and intercultural communication: linguistic analyses, language policies and didactic concepts*. Amsterdam: 249-264.
- Van Loey, A (1970): Schönfeld's Historische Grammatica van het Nederlands.