# Revealing Phonological Similarities between German and Dutch

*Karin Müller*

University of Amsterdam
Informatics Institute, Kruislaan 403, 1098 SJ Amsterdam, The Netherlands
kmueller@science.uva.nl

## Abstract

In this paper, we present an approach to automatically revealing phonological classes within historically related languages. A newly created bilingual German-Dutch pronunciation dictionary is used for learning phonological similarities between the onsets, nuclei and codas of these two languages via EM-based clustering. Our evaluation is twofold: we apply the models to predict from a German word the phonemes of a Dutch cognate. The results show that it is harder to predict the pronunciation of the nucleus and the coda than the onset. We also evaluate our approach qualitatively, finding meaningful classes caused by historical sound changes.

## 1. Introduction

German and Dutch are languages that exhibit a wide range of similarities. Beside similar syntactic features like word order and verb subcategorization frames, the languages share phonological features which are due to historical sound changes. These similarities are one reason why it is easier to learn a closely historically related language than languages from other language families: the learner's native language provides a valuable resource which can be used in learning the new language.

The knowledge about similarities on the lexical level is exploited in various fields. In machine translation, some approaches search for similar words (cognates) which are used to align parallel texts (e.g., [1]). *Technik-techniek* (technique) can be easily recognized as a cognate; recognizing *Pferd-paard* (horse), however, requires more knowledge about sound changes within the languages. The algorithms developed for machine translation search for similarities on the orthographic level, whereas some approaches to comparative and synchronic linguistics put their focus on similarities of phonological sequences. [2] gives an overview of current algorithms applied to the comparison of phonetic units. For instance, [3] computes the similarity of various language pairs, whereas [4] measure the phonetic distance between dialects. The above mentioned approaches strongly depend on parallel corpora which are time intensive and expensive to collect.

In our approach, we focus on generating automatically data which can be used as input to an unsupervised training procedure and with the aim of learning similar structures from these data using EM-based clustering. Our main assumption is that certain German-Dutch phoneme pairs from related stems occur more often and hence will appear in the same class with a higher probability than pairs not in related stems. The resulting classes should mirror the human ability to make use of similar phonological structures when learning related languages.

The paper is organized as follows: Section 2 presents related research. In Section 3, we describe the creation of our bilingual pronunciation dictionary which is used as input to the algorithm for automatically deriving phonological classes described in Section 4. In Section 5, we apply our classes to translating transcribed cognates and evaluate the results of this task. The second evaluation is presented in Section 6, where we interpret our best models. In Section 7, we discuss our results.

## 2. Previous Research

[5] analyze 800 Dutch and German cognates and present a theory of cross-linguistic phoneme correspondences. The list serves as evaluation corpus in our approach. They try to find phoneme correspondences in closely related languages. The results are tables of phonetic correspondences representing the counts for a certain German phoneme and their possible Dutch counterpart. [4] compared forty different Dutch dialects by measuring the phonetic difference of 101 pairs of words. They calculate a distance matrix which is subject to a heuristic clustering method. The work focuses on phonetic features and not on phonetic symbols like ours. [6] presents a method of discovering complex sound correspondences in bilingual word lists, using noisy word lists of Algonquian languages. The algorithm is evaluated against manually compiled sound correspondences. The algorithm incorporates knowledge about bilingual consonant and vowel pairs independent from phonotactic constraints. A translation task on the diachronic level is presented in [7], where he generates from a Proto-Slavic word a possible Polish version. Another generative model can be found in [8]. The model is applied to automatically translating Japanese words to English. The Japanese words – originally coming from English – are back transliterated to English using weighted finite-state transducers. In our approach, we aim at discovering similar correspondences between bilingual data.

## 3. Generation of a bilingual resource

In this section, we describe the resources used for our clustering algorithm. We take advantage of two on-line bilingual orthographic dictionaries[1] and the monolingual pronunciation dictionaries [9] in CELEX to automatically build a bilingual pronunciation dictionary.

In a first step, we extract from the German-Dutch orthographic dictionary 72,037 word pairs. Table 1 (1st subtable) displays a fragment of the extracted orthographic word pairs. Note, that we only allow one possible translation, namely the first one. Next, we automatically look up the pronunciation of the German and Dutch words in the monolingual part of CELEX. A word pair is considered for further analysis if the pronunciation of both words is found in CELEX. For instance, the first half of the word pair *Hausflur-huisgang* (corridor) does occur

[1] "http://deatch.de/niederlande/buch.htm"

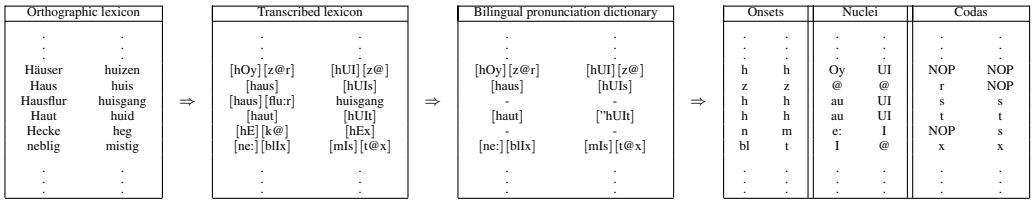


Table 1: Creation of the **German-Dutch input**: from a fragment of the orthographic lexicon (consisting of the words *houses, house, corridor, skin, hedge, misty*) ⇒ the automatically transcribed lexicon ⇒ the bilingual pronunciation dictionary ⇒ to the final bilingual onset, nucleus and coda lists ( left to right)

in the German part of CELEX but the second half is not contained within the Dutch part. Thus, this word pair is discarded. However, the words *Haus-huis* (house) are found in both monolingual pronunciation dictionaries and are used for further analysis. The automatic process of course introduces some noise to the dictionaries. We also find the word pair *neblig-mistig* (misty) which consists of two unrelated stems. Note that the transcription follows the CELEX conventions[2]. The result is a list of 44,415 transcribed German-Dutch word pairs. Figure 1(2nd subtable) shows the result of the look-up procedure. For instance, ["haus][3]-["hUIs] is the transcription of *Haus-huis* in the German-Dutch dictionary.

We aim at revealing phonological relationships between German-Dutch word pairs on the phonemic level, hence, we need something similar to an alignment procedure on the syllable level. Thus, we first extract only those word pairs which contain the same number of syllables. The underlying assumption is that words with a historically related stem often preserve their syllable structure. The only exception is that we do not use all inflectional paradigms of verbs to gain more data because they are often a reason for uneven syllable numbers (e.g., the past tense German suffix /tete/ is in Dutch /te/ or /de/). *Häuser-huizen* (houses) would be chosen both made up of two syllables; however, *Hecke-heg* (hedge) will be dismissed as the German word consists of two syllables whereas the Dutch word consists of one syllable. Figure 1 (3rd subtable) show the remaining items after this filtering process which represents the bilingual pronunciation dictionary. We split each syllable within the bilingual word lists into onset, nucleus and coda. All consonants left to the vowel are considered the onset and the consonants right to the vowel represent the coda. Empty onsets and codas are replaced by the word [NOP]. After this processing step, each word pair consists of the same number of onsets, nuclei and codas.

The final step is to extract a list of German-Dutch phoneme pairs. We can easily extract the bilingual onset, nucleus and coda pairs from the transcribed word pairs, shown in the fourth subtable of Figure 1. For instance, we extract the onset pair [h]-[h], the nucleus pair [au]-[UI] and the coda pair [s]-[s] from the German-Dutch word pair ["haus]-["hUIs]. With the described method, we obtain from the remaining 21,212 German-Dutch words, 59,819 German-Dutch onset, nucleus and coda pairs.

# 4. Phonological Clustering

In this section, we describe the unsupervised clustering method used for clustering of phonological units. Three- and five-dimensional EM-based clustering has been applied to monolingual phonological data [10] and two-dimensional clustering to syntax [11]. In our approach, we apply two-dimensional clustering to reveal classes of bilingual sound correspondences. The method is well-known but the application of probabilistic clustering to bilingual phonological data allows a new view on bilingual phonological processes. We choose EM-based clustering as we need a soft clustering technique which provides probabilities to deal with rather noisy input. The two main parts of EM-based clustering are (i) the induction of a smooth probability model over the data, and (ii) the automatic discovery of class structure in the data. We aim to derive a probability distribution $p(y)$ on bilingual phonological units $y$ from a large sample.

$$p(y) = \sum_{c \in C} p(c) \cdot p(y_{source}|c) \cdot p(y_{target}|c) \qquad (1)$$

The re-estimation formulas are given in [11] and our training regime dealing with the free parameters (e.g. the number of $|c|$ of classes) is described in Sections 4.1. The output of our clustering algorithm are classes with their class number, class probability and a list of class members with their probabilities.

| class 2 | 0.069 | |
|---|---|---|
| t | 0.633 | t 0.764 |
| ts | 0.144 | d 0.128 |
| s | 0.055 | |

The above table comes from our German-Dutch experiments and shows Class # 2 with its probability of 6.9%, the German onsets in the left column (e.g., [t] appears in this class with the probability of 63.3%, [ts] with 14.4% and [s] with 5.5%) and the Dutch onsets in the right column ([t] appears in this class with the probability of 76.4% and [d] with 12.8%). The examples presented in this paper are fragments of the full classes showing only those units with the highest probabilities.

### 4.1. Experiments with German-Dutch data

We use the 59,819 onset, nucleus and coda pairs as training material for our unsupervised training. Unsupervised methods require the variation of all free parameters to search for the optimal model. There are three different parameters which has to be varied: the initial start parameters, the number of classes and the number of re-estimation steps. Thus, we experiment with 10 different start parameters, 6 different numbers of classes (5, 10, 15, 20, 25 and 30[4]) and 20 steps of re-estimation. Our training regime yielded 1,200 onset, 1,200 coda and 1,000 nucleus models.

[2] "http: //www.ru.nl/celex/subsecs/section_doc.html"

[3] A syllable is transcribed within brackets ([syllable]).

[4] We did not experiment with 30 classes for nucleus pairs as there are fewer nucleus types than onset or coda types

## 5. Evaluation: Translation of cognates

We quantitatively evaluate our models with a translation task. The main idea is to take the transcription of a German word and to predict the most probable transcription of the Dutch cognate.

Hence, we extract 808 German-Dutch cognate pairs from a cognate database[5], consisting of 836 entries. As for the training data, we extract those pairs that consist of the same number of syllables because our current models are restricted to sound correspondences and can not discard complete syllables. We split our evaluation corpora into two parts serving as development database and as gold standard.

The task is then to predict the most probable translation of a German word to a Dutch word, e.g. the German word *durch* ([dUrx]) (through) should be translated to *door* ([do:r]) in Dutch. We evaluate all our onset, nucleus and coda models by calculating the most probable translation of the cognates from our development set and choosing the models with the highest onset, nucleus and coda precision separately. Only the best model (for onset, nucleus and coda prediction) is evaluated on the gold standard to avoid tuning to the development set. Using this procedure shows how our models perform on new data.

The table below shows the results of our best models by measuring the onset, nucleus and coda translation accuracy on our gold standard. We consider as baseline the number of cases where the German and the Dutch phonemes are the same.

| | Onset | Nucleus | Coda |
|---|---|---|---|
| accuracy | 80.7% | 50.7 % | 52.2 % |
| baseline | 40.6% | 42.7% | 49.2% |

The results show that the prediction of the onset is easier than predicting the nucleus or the coda. We achieve an onset accuracy of 80.7%. Although the set of possible nuclei is smaller than the set of onsets and codas, the prediction of the nuclei is much harder. The nucleus accuracy decreases to 50.7%. Codas seem to be slightly easier to predict than nuclei leading to a coda accuracy of 52.2%. Compared to the baseline, it seems that the onsets comprise less noise than the nucleus and the coda.

## 6. Evaluation: Interpretation of the Classes

In this section, we interpret our classes by manually identifying classes that show typical similarities between the two languages. Sometimes, the classes reflect sound changes in historically related stems. Our data is synchronic, and thus it is not possible to directly identify in our classes which sound changes took place (Modern German (G), and Modern Dutch (NL) did not develop from each other but from a common ancestor). However, we will try to connect the data to ancient languages such as Old High German (OHG), Middle High German (MHG), Middle Dutch (MNL), Old Dutch (ONL), Proto or West Germanic (PG, WG). Naturally, we can only go back in history as far as it is possible according to the information provided by the following literature: For Dutch, we use [12] and the online version of [13], and for German, [14]. We find that certain historic sound changes took place regularly, and thus, the results of these changes can be rediscovered in our synchronic classes. Figure 6 shows the historic relationship between West-Germanic languages. Naturally, a potential learner of a related language does not have to be aware of the historic links between languages but he/she can implicitly exploit the similarities such as the ones discovered in the classes.

[5] "http://www.itri.brighton.ac.uk/projects/metaphon/"

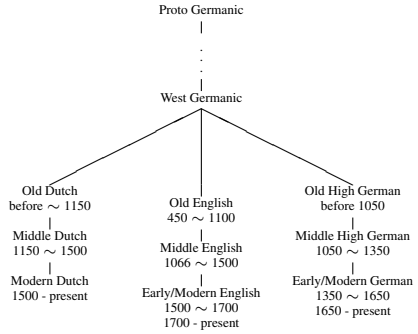


Figure 1: Family tree of West Germanic languages

The relationship of words from different languages can be caused by different processes: some words are simply borrowed from another language and adapted to a new language. Other language changes are due to phonology; e.g., the Proto Germanic word *muHs* was subject to diphthongization and changed to the German word *Maus* (MHG: mûs) and to the Dutch word *muis* (MNL: muus). On the synchonic level, we find [au] and [UI] in the German-Dutch models in the same class. There are also other phonological processes which apply to the nuclei, such as monophthongization, raising, lowering, backing and fronting. Other phonological processes can be observed in conjunction with consonants, such as assimilation, dissimilation, deletion and insertion. Some of the above mentioned phonological processes are the underlying processes of the subsequent described classes.

According to our evaluation presented in Section 5, the best onset model comprises 30 classes, the nucleus model 25 classes and the coda model 30 classes. We manually search for classes, which show interesting sound correspondences.

### 6.1. German-Dutch onset classes

| **class 20** 0.016 | | | |
|---|---|---|---|
| p | 0.747 | p | 0.902 |
| pf | 0.094 | x | 0.022 |
| r | 0.027 | | |
| x | 0.025 | | |
| f | 0.021 | | |

| **class 25** 0.012 | | | |
|---|---|---|---|
| S | 0.339 | sx | 0.189 |
| Sr | 0.172 | sxr | 0.162 |
| ts | 0.130 | s | 0.135 |
| tr | 0.122 | tr | 0.087 |
| z | 0.090 | st | 0.058 |

The German part of class # 20 reflects Grimm's first law which states that a West Germanic [p] is often realized as a [pf] in German. The underlying phonological process is that sounds are inserted in a certain context. The onsets of the Middle High German words *phat* (E: path) and *phert* (E: horse, L: paraverēredus) became the affricate [pf] in Modern German. In contrast to German, Dutch preserved the simple onsets from the original word form, as in *paard* (E: horse, MNL: peert) and *pad* (E: path, MNL: pat).

Class # 25 represents a class where the Dutch onsets are more complex than the onsets in German. From the Old High German word *scâf* (E: sheep) the onset /sc/ is assimilated in Modern German to [S] whereas the Dutch onset [sx] preserves the complex consonant cluster from the West Germanic word *skæpan* (E: sheep, MNL: scaep).

### 6.2. German-Dutch Nucleus classes

| **class 4** 0.054 | | | |
|---|---|---|---|
| U | 0.449 | O | 0.721 |
| O | 0.260 | U | 0.112 |
| Y | 0.079 | o: | 0.101857 |
| au | 0.072 | | |

We find in Class # 4 a lowering process. The German short high back vowel /U/ can be often transformed to the Dutch low

back vowel /O/. The underlying processes are that the Dutch vowel is sometimes lowered from /i/ to /O/; e.g., the Dutch word *gezond* (E: healthy, MNL: ghesont, WG: gezwind) comes from the West Germanic word *gezwind*. In Modern German, the same word changed to *gesund* (OHG: gisunt).

### 6.3. German-Dutch Coda classes

| **class 14** 0.027 | | | |
|---|---|---|---|
| m | 0.534 | m | 0.555 |
| n | 0.187 | NOP | 0.136 |
| NOP | 0.054 | x | 0.064 |
| mt | 0.042 | k | 0.06 |
| mst | 0.042 | mt | 0.055 |

| **class 23** 0.010 | | | |
|---|---|---|---|
| rt | 0.476 | rt | 0.521 |
| tst | 0.078 | t | 0.159 |
| rts | 0.068 | Nt | 0.049 |
| rst | 0.067 | lt | 0.029 |
| Nst | 0.047 | tst | 0.022 |
| | | st | 0.022 |

Class # 14 represents codas where plural and infinitive suffixes /en/, as in *Menschen-mensen* (E: humans) or *laufen-lopen* (E: to run), are reduced to a Schwa [@] in Dutch and thus appear in this class with an empty coda [NOP]. It also shows that certain German codas are assimilated by the alveolar sounds /d/ and /s/ from the original bilabial [m] to an apico-alveolar [n], as in *Boden* (E: ground, MHG: bodem) or in *Besen* (E: broom, MHG: bësem, OHG: pësamo). In Dutch, the words *bodem* (E: ground, MNL: bōdem, Greek: puthmēn), and *bezem* (E: broom, MNL: bēsem, WG: besman) kept the /m/.

Class # 23 comprises complex German codas which are less complex in Dutch. In the German word *Arzt* (E: doctor, MHG: arzât), the complex coda [tst] emerges. However in Modern Dutch, *arts* came from MNL *arst* or *arsate* (Latin: archiāter). We can also find the rule that German codas [Nst] of a 2nd person singular form of a verb are reduced to [Nt] in Dutch as in *bringst-brengt* (E: bring).

## 7. Discussion

We automatically generated a bilingual phonological corpus. The data is classified by using an EM-based clustering algorithm which is new in that respect that this method is applied to bilingual onset, nucleus and coda corpora. Revealing phonological relationships between languages is possible simply because the noisy data used comprises enough related words and syllables to learn from them the similar structure of the languages on the syllable-part level.

Our method differs from other approaches either in the comparison of different language pairs or in the different linguistic task. [5] is based on mere counts of phoneme correspondences; [6] works with bilingual phoneme correspondences (Algonquian data), although he worked on many language pairs [7], he did not compare German and Dutch; and [8] focus on the back-transliteration of Japanese words to English. Thus, we regard our approach as a thematic complement and not as an overlap to former approaches.

Naturally, our method depends on the resources. That means that we can only learn those phoneme correspondences which are available in our data. Thus, metathesis which applies to onsets and codas can not be directly observed as the syllable parts are modeled separately. In the Dutch word *borst* (E: breast, ONL: bructe), the /r/ shifted from the onset to the coda whereas in German it stayed in the onset (G: Brust). We also rely on the CELEX builders, who followed different transcription strategies for the German and Dutch parts. For instance, elisions occur in the Dutch lexicon but not in the German part. In *luchtdruk* (E: air pressure) ["lUG][drUk], the coda consonant /t/ disappears in the Dutch word but not in the German word *Luftdruck*.

The results on predicting the phonemes of cognates might be improved by increasing the size of the databases. An interesting point for future work is to apply the methods for the identification of cognates to the bilingual word-list similar to the work done by [7]. Beyond the increase in data, a great challenge is to develop models that can express the sound change on the diachronic level adumbrated in Section 6.

We showed that onsets are easier to predict than nuclei or codas which points out that onsets across the two languages are more stable than nuclei or codas. We also believe that the results of our experiments – in particular, the classes presenting probable sound correspondences – might be useful for language learning. If the classes are augmented by exemplifying word pairs, a language learner can use the classes for more systematic learning.

## 8. References


[1] M. Simard, G. F. Foster, and P. Isabelle, "Using cognates to align sentences in bilingual corpora," in *Proceedings of TMI-92*, Montreal Canada, 1992.

[2] B. Kessler, "Phonetic comparison algorithms," *Transactions of the Philological Society*, 2005, in press.

[3] G. Kondrak, "A New Algorithm for the Alignment of Phonetic Sequences," in *Proceedings of NAACL 2000*, Seattle, WA, 2000.

[4] J. Nerbonne and W. Heeringa, "Measuring Dialect Distance Phonetically," in *Proceedings of the third meeting of the SIGPHON at ACL*, 1997, pp. 11–18.

[5] L. Cahill and C. Tiberius, "Cross-linguistic phoneme correspondences," in *Proceedings of ACL 2002*, Taipai, Taiwan, 2002.

[6] G. Kondrak, "Identifying Complex Sound Correspondences in Bilingual Wordlists," in *Proceedings of CICLING 2003*, Mexico City, 2003.

[7] ——, "Algorithms for Language Reconstruction," Ph.D. dissertation, University of Toronto, 2002.

[8] K. Knight and J. Graehl, "Machine transliteration." *Computational Linguistics*, vol. 24, no. 4, pp. 599–612, 1998.

[9] H. R. Baayen, R. Piepenbrock, and H. van Rijn, "The CELEX lexical database—Dutch, English, German," (Release 1)[CD-ROM]. Philadelphia, PA: Linguistic Data Consortium, Univ. Pennsylvania, 1993.

[10] K. Müller, B. Möbius, and D. Prescher, "Inducing Probabilistic Syllable Classes Using Multivariate Clustering," in *Proc. 38th Annual Meeting of the ACL*, Hongkong, China, 2000.

[11] M. Rooth, S. Riezler, D. Prescher, G. Carroll, and F. Beil, "Inducing a Semantically Annotated Lexicon via EM-Based Clustering," in *Proc. 37th Annual Meeting of the ACL*, College Park, MD, 1999.

[12] J. de Vries, *Nederlands Etymologisch Woordenboek*. Leiden: Brill, 1997.

[13] M. Philippa, F. Debrabandere, and A. Quak, *Etymologisch Woordenboek van het Nederlands deel 1: A t/m E.* Amsterdam: Amsterdam University Press, 2004, vol. 1, "http://www.etymologie.nl/".

[14] T. Burch, J. Fournier, and K. Gärtner, "Mittelhochdeutsche Wörterbücher auf CD-ROM und im Internet ," *Akademie-Journal*, vol. 2, pp. 17–24, 1998, online Wörterbuch, "http://www.mwv.uni-trier.de/index.html".