Revealing Phonological Similarities between Related Languages from Automatically Generated Parallel Corpora

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

Abstract

In this paper, we present an approach to automatically revealing phonological correspondences within historically related languages. We create two bilingual pronunciation dictionaries for the language pairs German-Dutch and German-English. The data is used for automatically learning phonological similarities between the two language pairs via EM-based clustering. We apply our models to predict from a phonological German word the phonemes of a Dutch and an English cognate. The similarity scores show that German and Dutch phonemes are more similar than German and English phonemes, which supplies statistical evidence of the common knowledge that German is more closely related to Dutch than to English. We assess our approach qualitatively, finding meaningful classes caused by historical sound changes. The classes can be used for language learning.

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. Although English also belongs to the West Germanic languages, German and Dutch share more lexical entries with a common root than German and English.

The knowledge about language 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., Simard et al. (1992)). The word triple *Text-tekst-text* ([tEkst] in German, Dutch and English) can be easily recognized as a cognate; recognizing *Pfeffer-peper-pepper* ([pfE][f@r]-[pe:][p@r])-[pE][p@r*]), 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. Covington (1996), for instance, suggests different algorithms to align the phonetic representation of words of historical languages. Kondrak (2000) presents an algorithm to align phonetic sequences by computing the similarities of these words. Nerbonne and Heeringa (1997) use phonetic transcriptions to measure the phonetic distance between different dialects. The above mentioned approaches presuppose either parallel texts of different languages for machine translation or manually compiled lists of transcribed cognates/words for analyzing synchronic or diachronic word pairs. Unfortunately, transcribed bilingual data are scarce and it

Proceedings of the ACL Workshop on Building and Using Parallel Texts, pages 33–40,
Ann Arbor, June 2005. ©Association for Computational Linguistics, 2005

is labor-intensive to collect these kind of corpora. Thus, we aim at exploiting electronic pronunciation dictionaries to overcome the lack of data.

In our approach, we automatically generate data as input to an unsupervised training regime and with the aim of automatically learning similar structures from these data using Expectation Maximization (EM) based clustering. Although the generation of our data introduces some noise, we expect that our method is able to automatically learn meaningful sound correspondences from a large amount of data. Our main assumption is that certain German/Dutch and German/English 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. We assume that the historical sound changes are hidden information in the classes.

The paper is organized as follows: Section 2 presents related research. In Section 3, we describe the creation of our bilingual pronunciation dictionaries. The outcome is used as input to the algorithm for automatically deriving phonological classes described in Section 4. In Section 5, we apply our classes to a transcribed cognate list and measure the similarity between the two language pairs. A qualitative evaluation is presented in Section 6, where we interpret our best models. In Sections 7 and 8, we discuss our results and draw some final conclusions.

2 Previous Research

Some approaches to revealing sound correspondences require clean data whereas other methods can deal with noisy input. Cahill and Tiberius (2002) use a manually compiled cognate list of Dutch, English and German cognates and extract cross-linguistic phoneme correspondences. The results[1] contain the counts of a certain German phoneme and their possible English and Dutch counterparts. The method presented in Kondrak (2003), however, can deal with noisy bilingual word lists. He generates sound correspondences of various Algonquian languages. His algorithm considers them as possible candidates if their likelihood scores lie above a certain minimum-strength threshold. The candidates are evaluated against manually compiled sound correspondences. The algorithm is able to judge whether a bilingual phoneme pair is a possible sound correspondence. Another interesting generative model can be found in Knight and Graehl (1998). They train weighted finite-state transducers with the EM algorithm which are applied to automatically transliterating Japanese words - originated from English - back to English. In our approach, we aim at discovering similar correspondences between bilingual data represented in the classes. The classes can be used to assess how likely a bilingual sound correspondence is.

3 Generation of two parallel Corpora

In this section, we describe the resources used for our clustering algorithm. We take advantage of two on-line bilingual orthographic dictionaries[2] and the monolingual pronunciation dictionaries (Baayen et al., 1993) in CELEX to automatically build two bilingual pronunciation dictionaries.

In a first step, we extract from the German-Dutch orthographic dictionary 72,037 word pairs and from the German-English dictionary 155,317. Figures 1 and 2 (1st table) display a fragment of the extracted orthographic word pairs. Note that we only allow one possible translation, namely the first one.

In a next step, we automatically look up the pronunciation of the German, Dutch and English 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 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 all three monolingual pronunciation dictionaries and are used for further analysis. Note that the transcription and syllabification of the words are defined in CELEX.

The result is a list of 44,415 transcribed German-Dutch word pairs and a list of 63,297 transcribed German-English word pairs. Figures 1 and 2 (2nd table) show the result of the look-up procedure. For instance, [”haus][3]-[”hUIs] is the transcription of *Haus-huis* in the German-Dutch dictionary, while

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

[2] http://deatch.de/niederlande/buch.htm
http://branchenportal-deutschland.aus-stade.de/englisch-deutsch.html

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

Orthographic lexicon	
⋮	⋮
Häuser	huizen
Haus	huis
Hausflur	huisgang
Haut	huid
Hautarzt	huidarts
⋮	⋮

⇒

Transcribed lexicon	
⋮	⋮
[”hOy] [z@r]	[hUI] [z@]
[”haus]	[”hUIs]
[”haus] [flu:r]	huisgang
[”haut]	[”hUIt]
[haut] [”a:rtst]	[hUId] [Arts]
⋮	⋮

⇒

Bilingual pronunciation dictionary	
⋮	⋮
[”hOy] [z@r]	[hUI] [z@]
[”haus]	[”hUIs]
-	-
[”haut]	[”hUIt]
[haut] [”a:rtst]	[hUId] [Arts]
⋮	⋮

⇒

Onsets		Nuclei		Codas	
⋮	⋮	⋮	⋮	⋮	⋮
h	h	Oy	UI	NOP	NOP
z	z	@	@	r	NOP
h	h	au	UI	s	s
h	h	au	UI	t	t
h	h	au	UI	t	d
NOP	NOP	a:	A	rtst	rts
⋮	⋮	⋮	⋮	⋮	⋮

Figure 1: Creation of the **German-Dutch input**: from the orthographic lexicon - the automatically transcribed lexicon - the bilingual dictionary - to the final bilingual onset, nucleus and coda lists (left to right)

Orthographic lexicon	
⋮	⋮
Häuser	houses
Haus	house
Hausflur	corridor
Haut	skin
Hautarzt	dermatologist
⋮	⋮

⇒

Transcribed lexicon	
⋮	⋮
[”hOy] [z@r]	[”haU] [zIz]
[”haus]	[haUs]
[”haus] [flu:r]	[”kO] [rI] [dO:r⋆]
[”haut]	[skIn]
[haut] [”a:rtst]	[d3:] [m@] [”tO]- -[l@] [dZIst]
⋮	⋮

⇒

Bilingual pronunciation dictionary	
⋮	⋮
[”hOy] [z@r]	[”haU] [zIz]
[”haus]	[haUs]
-	-
[”haut]	[skIn]
-	-
⋮	⋮

⇒

Onsets		Nuclei		Codas	
⋮	⋮	⋮	⋮	⋮	⋮
h	h	Oy	aU	NOP	NOP
z	z	@	I	r	z
h	h	au	aU	s	s
h	sk	au	I	t	n
⋮	⋮	⋮	⋮	⋮	⋮

Figure 2: Creation of the **German-English input**: from the orthographic lexicon - the automatically transcribed lexicon - the bilingual dictionary - to the final bilingual onset, nucleus and coda lists (left to right)

[”haus]-[haUs] is the transcription of *Haus-house* in the German-English part.

We aim at revealing phonological relationships between German-Dutch and German-English 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/). *Hautarzt-huidarts* would be chosen both made up of two syllables; however, *Hautarzt-dermatologist* will be dismissed as the German word consists of two syllables whereas the English word comprises five syllables. Figures 1 and 2 (3rd table) show the remaining items after this filtering process. We split each syllable within the bilingual word lists into onset, nucleus and coda. All consonants to the left of the vowel are considered the onset. The consonants to the right of 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 and German-English phoneme pairs. It is easy to extract the bilingual onset, nucleus and coda pairs from the transcribed word pairs (fourth table of Figures 1 and 2). 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 and 13,067 German-English words, 59,819 German-Dutch and 35,847 German-English 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 (Müller et al., 2000) and two-dimensional clustering to syntax (Rooth et al., 1999). 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 technique which provides probabilities to deal with noise in the training data. 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 ($p(c)$ denotes the class probability, $p(y_{source}|c)$ is the probability of a phoneme of the source language given class c, and $p(y_{target}|c)$ is the probability of a phoneme of the target language given class c).

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

The re-estimation formulas are given in (Rooth et al., 1999) and our training regime dealing with the free parameters (e.g. the number of $|c|$ of classes) is described in Sections 4.1 and 4.2. 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 have 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 yields 1,200 onset, 1,200 coda and 1,000 nucleus models.

4.2 Experiments with German-English data

Our training material is slightly smaller for German-English than for German-Dutch. We derive 35,847 onset, nucleus and coda pairs for training. The reduced training set is due to the structure of words which is less similar for German-English words than for German-Dutch words leading to words with unequal syllable numbers. We used the same training regime as in Section 4.1, yielding the same number of models.

5 Similarity scores of the syllable parts

We apply our models to a translation task. The main idea is to take a German phoneme and to predict the most probable Dutch and English counterpart.

Hence, we extract 808 German-Dutch and 738 German-English 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 do not allow the deletion of syllables. We split our corpus into two parts by putting the words with an even line number in the development database and the words with an uneven line number in the gold standard database. The development set and the gold standard corpus consist of 404 transcribed words for the German to Dutch translation task and of 369 transcribed words for the German to English translation task.

The task is then to predict the translation of German onsets to Dutch onsets taken from German-Dutch cognate pairs, e.g. the models should predict from the German word *durch* ([dUrx]) (through), the Dutch word *door* ([do:r]). If the phoneme correspondence, [d]:[d], is predicted, the similarity score of the onset model increases. The nucleus score increases if the nucleus model predicts [U]:[o:] and the coda score increases if the coda model predicts [rx]:[r]. We assess all our onset, nucleus and coda models

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

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

German to Dutch			German to English		
Onset	Nucleus	Coda	Onset	Nucleus	Coda
80.7%	50.7 %	52.2 %	69.6%	17.1%	28.7%

Table 1: Similarity scores for syllable parts of cognates indicating that German is closer related to Dutch than to English.

by measuring the most probable phoneme translations of the cognates from our development set. We choose the models with the highest onset, nucleus and coda scores. Only the models with the highest scores (for onset, nucleus and coda prediction) are applied to the gold standard to avoid tuning to the development set. Using this procedure shows how our models perform on new data. We apply our scoring procedure to both language pairs.

Table 1 shows the results of our best models by measuring the onset, nucleus and coda translation scores on our gold standard. The results point out that the prediction of the onset is easier than predicting the nucleus or the coda. We achieve an onset similarity score of 80.7% for the German to Dutch task and 69.6% for the German to English task. 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 similarity score decreases to 50.7% and to 17.1% for German-English respectively. Codas seem to be slightly easier to predict than nuclei leading to a coda similarity score of 52.2% for German-Dutch and to 28.7% for German-English.

The comparison of the similarity scores from the translation tasks of the two language pairs indicates that predicting the phonological correspondences from German to Dutch is much easier than from German to English. These results supply statistical evidence that German is historically more closely related to Dutch than to English. We do not believe that the difference in the similarity scores are due to the different size of the training corpora but rather to their closer relatedness. Revealing phonological relationships between languages is possible simply because the noisy training data comprise enough related words to learn from them the similar structure of the languages on the syllable-part level.

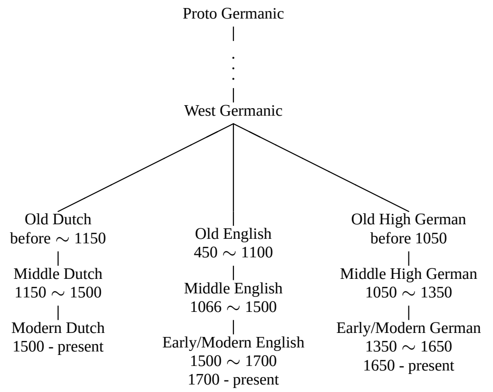

Figure 3: Family tree of West Germanic languages

6 Evaluation: Interpretation of the Classes

In this section, we interpret our classes by manually identifying classes that show typical similarities between the two language pairs. 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), Modern English (E) 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), Old English (OE), 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 de Vries (1997) and the online version of Philippa et al. (2004), for English, an etymological dictionary (Harper, 2001) and for German, Burch et al. (1998). 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 3 shows the historic relationship between the three languages. 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.

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. *Papagei-papegaai* (parrot) is borrowed from Arabic and adapted to German and Dutch phonetics, where the /g/ is pronounced in German as a voiced velar plosive and in Dutch as an unvoiced velar fricative.

Other language changes are due to phonology; e.g., the Old English word [mus] (PG: muHs) was subject to diphthongization and changed to *mouse* ([maUs]) in Modern English. A similar process took place in German and Dutch, where the same word changed to the German word *Maus* (MHG: mûs) and to the Dutch word *muis* (MNL: muus). On the synchronic level, we find [au] and [aU] in the same class of a German-English model and [au] and [UI] in a German-Dutch model. 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.

6.1 German-Dutch classes

According to our similarity scores 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.

6.1.1 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		

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	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

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).

6.1.2 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).

6.1.3 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 # 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	0.010		
rt	0.476	rt	0.521
tst	0.0782	t	0.159
rts	0.068	Nt	0.049
rst	0.067	lt	0.029
Nst	0.047	tst	0.022
t	0.023	rd	0.022
rtst	0.022	st	0.022
kt	0.021	rts	0.021
		xt	0.021

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).

6.2 German-English classes

The best German-English models contain 30 onset classes, 20 nucleus classes, and 10 coda classes. Our German-English models are noisier than the German-Dutch ones, which again points at the closer relation between the German and Dutch lexicon. However, when we analyze the 30 onset classes, we find meaningful processes as for German-Dutch.

6.2.1 Onset classes

class 23	0.016		
f	0.720	f	0.648
Sp	0.105	sp	0.131
z	0.044	v	0.059
S	0.012		
v	0.011		
...			
Spr	0.005		
sp	0.003		

Class # 23 shows that a complex German onset [Spr] preserves the consonant cluster, as in *sprechen* (E: to speak, OHG: sprehhan, PG: sprekanan). Modern English, however, deleted the /r/ to [sp], as in *speak* (OE: sprecan). Another regularity can be found: the palato-alveolar [S] in the German onset [Sp] is realized in English as the alveolar [s] in [sp]. Both the German word *spinnen* and the English word *spin* come from *spinnan* (OHG, OE).

class 3	0.051		
z	0.489	s	0.617
ts	0.170	z	0.143
s	0.087		

Class # 3 displays the rule that in many loan words, the onset /c/ is realized in German as [ts] and in English as [s] in *Akzent-accent* (Latin: accentus).

6.2.2 Nucleus classes

class 8	0.044		
o:	0.449	@U	0.425
y:	0.123	@	0.201
ai	0.055	O	0.115
		u:	0.048

In some loan words, we find that an original /u/ or /o/ becomes in German the long vowel [o:] and in English the diphthong [@U], as in *Sofa-sofa* (Arabic: suffah) or in *Foto-photo* (Latin: Phosphorus). The diphthongization in English usually applies to open syllables with the nucleus /o/, as shown in class # 8.

6.2.3 Coda classes

Class # 6 displays the present participle suffix /end/, which is realized in English as /ing/ (OE: -ende), as in *backend-baking*.

class 6	0.056		
nt	0.707	N	0.846
N	0.075	NOP	0.072
lnt	0.058	nt	0.041
NOP	0.049	v	0.009
rnt	0.047	s	0.008

7 Discussion

We automatically generated two bilingual phonological corpora. 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. The method provides a probability model over bilingual syllable parts which is exploited to measure the similarity between the language pairs German-Dutch and German-English. The method is able to generalize from the data and reduces the noise introduced by the automatic generation process. Highly probable sound correspondences appear in very likely classes with a high probability whereas unlikely sound correspondences receive lower probabilities.

Our approach differs from other approaches either in the method used or in the different linguistic task. Cahill and Tiberius (2002) is based on mere counts of phoneme correspondences; Kondrak (2003) generates Algonquian phoneme correspondences which are possible according to his translation models; Kondrak (2004) measures if two words are possible cognates; and Knight and Graehl (1998) 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.

The presented approach depends on the available resources. That means that we can only learn those phoneme correspondences which are represented in the bilingual 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* (ONL: bructe), the /r/ shifted from the onset to the coda whereas in English and German (*breast-Brust*), it remained in the onset. We are also dependent 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. The coda consonant /t/ in *lucht* (air) disappears in the Dutch word *luchtdruk* (E: air pressure), [”lUG][drUk], but not in the German word *Luftdruck*, [lUft][drUk].

We assume that the similarity scores of the syllable parts might be sharpened by increasing the size of the databases. A first possibility is to take the first transcribed translation and not the first translation in general. As often the first translation is not contained in the pronunciation dictionary.

Our current data generation process also introduces unrelated word pairs such as *Haut-skin* ([haut]-[skIn]). However, it is very unlikely that related words do not include similar phonemes. Thus, this word pair should be excluded. Exploiting this knowledge could lead to cleaner input data.

8 Conclusions and Future Work

We presented a method to automatically build bilingual pronunciation dictionaries that can be used to reveal phonological similarities between related languages. In general, our similarity scores show that the lexicons of German and Dutch are closer related than German and English. Beside the findings about the relatedness between the two language pairs, we think that the classes might be useful for language learning. An interesting point for future work is to apply the methods developed for the identification of cognates to our bilingual word-lists. Beyond the increase in data, a great challenge is to develop models that can express sound changes on the diachronic level adumbrated in Section 6. We also believe that a slightly modified version of our method can be applied to other related language pairs by using the transcription of morphemes.

9 Acknowledgments

This research was supported by the Netherlands Organization for Scientific Research under project nr. 220-80-001. I am especially endebted to B. Möbius and G. Dogil for their support and comments during a research visit in Stuttgart, as well as to D. Ahn, D. Prescher and E. Tjong Kim Sang for comments.

References

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

Thomas Burch, Johannes Fournier, and Kurt Gärtner. 1998. Mittelhochdeutsche Wörterbücher auf CD-ROM und im Internet . *Akademie-Journal*, 2:17–24. ”http://www.mwv.uni-trier.de/index.html”.

Lynne Cahill and Carole Tiberius. 2002. Cross-linguistic phoneme correspondences. In *Proceedings of ACL 2002*, Taipai, Taiwan.

Michael A. Covington. 1996. An Algorithm to Align Words for Historical Comparison. *Computational Linguistics*, 22(4):481–496.

Jan de Vries. 1997. *Nederlands Etymologisch Woordenboek*. Brill, Leiden.

Daniel Harper. 2001. Online Etymology Dictionary. ”http://www.etymonline.com”.

Kevin Knight and Jonathan Graehl. 1998. Machine transliteration. *Computational Linguistics*, 24(4):599–612.

Grzegorz Kondrak. 2000. A New Algorithm for the Alignment of Phonetic Sequences. In *Proceedings of NAACL 2000*, Seattle, WA.

Grzegorz Kondrak. 2003. Identifying Complex Sound Correspondences in Bilingual Wordlists. In *Proceedings of CICLING 2003*, Mexico City.

Grzegorz Kondrak. 2004. Combining evidence in cognate identification. In *Proceedings of Canadian AI 2004*, pages 44–59.

Karin Müller, Bernd Möbius, and Detlef Prescher. 2000. Inducing Probabilistic Syllable Classes Using Multivariate Clustering. In *Proc. 38th Annual Meeting of the ACL*, Hongkong, China.

John Nerbonne and Wilbert Heeringa. 1997. Measuring Dialect Distance Phonetically. In *Proceedings of the third meeting of the SIGPHON at ACL*, pages 11–18.

Marlies Philippa, Frans Debrabandere, and Arend Quak. 2004. *Etymologisch Woordenboek van het Nederlands deel 1: A t/m E*, volume 1. Amsterdam University Press, Amsterdam. ”http://www.etymologie.nl/”.

Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. 1999. Inducing a Semantically Annotated Lexicon via EM-Based Clustering. In *Proc. 37th Annual Meeting of the ACL*, College Park, MD.

Michel Simard, George F. Foster, and Pierre Isabelle. 1992. Using cognates to align sentences in bilingual corpora. In *Proceedings of TMI-92*, Montreal Canada.