Chapter 14
Lexical Modeling for Proper name Recognition in Autonomata Too

Bert Réveil, Jean-Pierre Martens, Henk van den Heuvel, Gerrit Bloothooft, and Marijn Schraagen

14.1 Introduction

Points of Interest business applications are strongly emerging on the ICT market, in particular in high-end navigation systems. For cars for instance, there is a high safety issue, and voice-driven navigation systems are appealing because they offer hands- and (partly) eye-free operation. However, Points of Interest like company names, hotel and restaurant names, names of attraction parks and museums, etc., often contain non-native parts. Moreover, the targeted application must be usable by non-native as well as native speakers. This means that there are considerable cross-lingual effects to cope with, which implies that the challenges for the automatic recogniser are high. At the start of the project (February, 2008) there was indeed substantial evidence [1–8] that state-of-the-art ASR technology was not yet good enough to enable a sufficiently reliable voice-driven POI business service.

The general project aim was therefore to improve name recognition accuracy by better coping with the large degree of variations observed in the POI pronunciations. The specific aim was to improve the recognition of (1) native Dutch/Flemish pronunciations of Dutch/Flemish POI, (2) native Dutch/Flemish pronunciations of foreign POI, and (3) non-native pronunciations of Dutch and Flemish POI. An important constraint was that the envisaged approach would have to be easily transferable from one application domain (e.g. car navigation) to another (e.g.

B. Réveil (✉) · J.-P. Martens
Ghent University, ELIS-DSSP, Sint-Pietersnieuwstraat 41, 9000 Gent, Belgium
e-mail: bert.reveil@elis.ugent.be; martens@elis.ugent.be

H. van den Heuvel
Radboud University, CLST, Erasmusplein 1, 6500 HD Nijmegen, The Netherlands
e-mail: H.vandenHeuvel@let.ru.nl

G. Bloothooft · M. Schraagen
Utrecht Institute of Linguistics, OTS, Trans 10, 3512 JK Utrecht, The Netherlands
e-mail: g.bloothooft@uu.nl

P. Spyns and J. Odijk (eds.), *Essential Speech and Language Technology for Dutch*, Theory and Applications of Natural Language Processing,
DOI 10.1007/978-3-642-30910-6_14,

telephone-based services for ordering medication, whiskey brands, etc.). Therefore, we contemplated an approach that would require no transcribed spoken name utterances from the targeted application domain. The domain knowledge would have to be provided in the form of example phonemic name transcriptions that are easy to acquire from people who know the application domain. The method would initially be developed and assessed for the domain of person name and geographical name recognition because for this domain transcribed utterances were available for development and evaluation, thanks to the Autonomata project (cf. Chap. 4, p. 61 on Autonomata resources). Subsequently, it would be transferred to the domain of POI recognition for which no transcribed development data were available yet.

The general concept of our methodology is that new pronunciation variants are generated with so-called P2P converters that apply automatically learned context-dependent transformation rules on the output of general-purpose G2P converters.

The rest of this chapter is organised as follows. In Sect. 14.2 we give a survey of multilingual pronunciation and acoustic modeling methods that were previously proposed for improving proper name recognition. In that section we also discuss the experiments that we conducted in order to define a state-of-the-art baseline system. In Sect. 14.3 we try to quantify how much further improvement is possible by means of more advanced pronunciation modeling techniques. In Sect. 14.4 we discuss the approach that we developed and the elements that make it unique. In Sect. 14.5 we offer an experimental validation of our method in the person and geographical name domains as well as in the targeted POI domain. The main conclusions of our work are formulated in Sect. 14.6.

14.2 Formerly Proposed Approaches

It has been shown by many authors that when cross-lingual factors come into play both acoustic and lexical modeling techniques can help to improve the ASR accuracy. For proper name recognition this is evidently the case, which is why we briefly review some of these techniques and why we assessed them when applied to proper name recognition.

14.2.1 Acoustic Modeling Approaches

Acoustic modeling tries to cope with the different ways in which an intended sound (a phoneme) can be articulated by the speaker. For the particular case of accented speech, a well known recipe to improve the recognition is to collect a small accented speech corpus and to adapt native acoustic models to the considered accent on the basis of this corpus. Popular adaptation methods in this respect are maximum likelihood linear regression (MLLR) [9] and maximum a posteriori (MAP) adaptation [10]. In [11], this technique yielded a 25 % improvement for the recognition of

English text spoken by Japanese natives with a low-proficiency in English. In [12], MLLR and MAP adaptation were used sequentially to adapt context-independent native acoustic models to an a priori known accent. Improvements of over 50 % could be attained in the context of an automated vocal command system.

An alternative approach is to start with a multilingual phoneme set and multilingual training data and to train context-dependent phoneme models on data from all languages in which the corresponding phonemes appear. By doing so for a bilingual set-up (German as native and English as foreign language), [13] could improve the recognition of (partly) English movie titles read by German natives by 25 % relative. In [14], the problem of recognising accented English speech embedded in Mandarin speech is tackled. Improvements of around 20 % relative over the standard multilingual approach were obtained by merging the output distribution of each bilingual model state with that of a related Mandarin accented English model state. The related state is identified automatically using a measure of the acoustic distance between states.

In [15,16], the more challenging case of multiple foreign accents was considered. French commands and expressions uttered by speakers from 24 different countries were recognised using a baseline French system, and a multilingual system that was obtained by supplementing the French acoustic models with three foreign (English, German and Spanish) acoustic model sets that were trained on speech from the corresponding languages. The multilingual acoustic models did improve the recognition for English and Spanish speakers (by about 15–20 %), but unexpectedly, degraded it for German speakers (by about 25 %). Furthermore, there was also a significant degradation for native French speakers and non-native French speakers of non-modeled languages.

14.2.2 Lexical Modeling Approaches

Lexical modeling deals with the phonetisation process, defined as the internal conversion of the orthography to a phonemic transcription that then serves as the basis for the articulation. It is generally known that non-native speakers often perform a non-standard phonetisation. In order to deal with this phenomenon, lexical modeling tries to enrich a baseline lexicon with the most frequently occurring non-standard phonetisations. One popular recipe is to add transcriptions emerging from G2P converters that implement the phonetisation rules of the most relevant foreign languages. In [1], Dutch, English and French G2P transcriptions were included for all entries (about 500) in a pronunciation dictionary containing Dutch, English, French and other names. Using optimised language dependent weights for the transcriptions, the name error rate could be reduced by about 40 % for native Dutch speakers, 70 % for French speakers, 45 % for English speakers and 10 % for other foreign speakers.

A similar approach was adopted in [6], but in a larger scale set-up with a vocabulary of 44K person names that occur in the US. Two baseline pronunciation

dictionaries were constructed: one with handcrafted typical native US English transcriptions (TY) and one with transcriptions emerging from a native US English G2P converter. Then, new variants were generated by eight foreign G2P converters covering all foreign language origins of the names occurring in the data set. Using n-gram grapheme models as language identifiers, likelihoods for the name source languages were computed and the transcriptions generated by the top two foreign G2P converters were added to the baseline lexicons. The variants caused a 25 % reduction of the name error rate for all names uttered by non-native speakers, irrespective of the baseline lexicon. However, the error rate reduction was only 10 % for the native utterances of foreign names and insignificant for the native utterances of native names.

14.2.3 Assessment of Established Approaches

In order to assess the formerly presented approaches, we performed recognition experiments with the Dutch version of the commercially available state-of-the-art Nuance VoCon 3200 engine.[1] The engine was a black box for us, but nevertheless it permitted us to investigate some of the proposed recipes as it was delivered with two acoustic models:

- AC-MONO: a monolingual acoustic model that was trained on native Dutch speech. The underlying phoneme set consists of 45 phonemes.
- AC-MULTI: a multilingual acoustic model that was trained on the same Dutch speech, but supplemented with equally large amounts of UK English, French and German speech. The underlying phoneme set consists of 80 phonemes and models of phonemes appearing in multiple languages have thus seen data from all these languages.

Experiments were conducted on the Autonomata Spoken Name Corpus (ASNC).[2] This corpus contains isolated proper name utterances from 240 speakers, and each speaker has read 181 names (person names and geographical names). The *speaker tongue*, defined as the mother tongue of the speaker, and the *name source*, defined as the language of origin of the name, in the ASNC is either Dutch, English, French, Turkish or Moroccan Arabic. In what follows, we have split the corpus into cells on the basis of these variables. The cell (DU,EN) for instance, contains the recordings of Dutch speakers reading English names. A division in training and test data (70–30 %) was made in such a way that any overlap between speakers and names in the two sets was avoided. In the present chapter, the training set is only used to provide phonemic transcriptions for examplary names that do not occur in the test set. No knowledge about the speech recordings, through e.g.

[1] www.nuance.com/for-business/by-product/automotive-products-services/vocon3200/index.htm

[2] For a detailed corpus description, we refer the reader to Chap. 4 of this book, Sect. 4.2, p. 62.

Table 14.1 Number of tokens per (speaker tongue, name source) combination in the ASNC test set

(DU,DU)	(DU,EN)	(DU,FR)	(DU,NN2)	(DU,ALL)
4,440	851	414	992	6,697
(DU,DU)	(EN,DU)	(FR,DU)	(NN2,DU)	(ALL,DU)
4,440	1,800	720	2,280	9,240

auditorily verified transcriptions, is employed (in contrast to [17], where we did use that information).

For the interpretation of results, a distinction was made between the native language (Dutch), non-native languages most native speakers speak/understand to some extent (English and French, called NN1 languages), and non-native languages most speakers are not familiar with at all (Turkish and Moroccan Arabic). The latter two languages are always pooled to form one 'language' called NN2. Table 14.1 shows the number of test set utterances in the different (speaker tongue, name source) cells of interest.

We chose to employ the Name Error Rate (NER) as our evaluation metric. It is defined as the percentage of name utterances that are not correctly recognised.

Figure 14.1 shows how the NER in the considered cells is affected by (1) decoding the utterances with a monolingual/multilingual acoustic model, (2) including foreign G2P transcriptions in the lexicon, and (3) adopting a monolingual/multilingual phoneme set in the lexicon.[3] The recognition vocabulary consists of all the 3,540 unique names appearing in the ASNC.

The three most important conclusions that can be drawn from the figure are the following:

1. Supplementing the lexicon with transcriptions emerging from a non-native G2P converter helps a lot for the recognition of non-native names originating from the corresponding language (the English/French transcriptions were only added for the French/English names).
2. Replacing a monolingual by a multilingual acoustic model significantly raises the recognition accuracy for non-native speakers reading native names, at least as long as the non-native language under concern was included in the acoustic model training data.
3. Nativising the non-native G2P transcriptions does not (significantly) reduce the gains that can be achieved with a multilingual acoustic model.

The first two conclusions confirm the formerly cited observations and the fact that in the target applications, the two techniques act complementary. The last conclusion

[3] A monolingual phoneme set implies that we need *nativised* Dutch versions of the foreign G2P transcriptions. These were obtained by means of a manual mapping of the foreign phonemes onto the Dutch phoneme set. The mapping was based on our own linguistic intuition, without prior knowledge of the recordings.

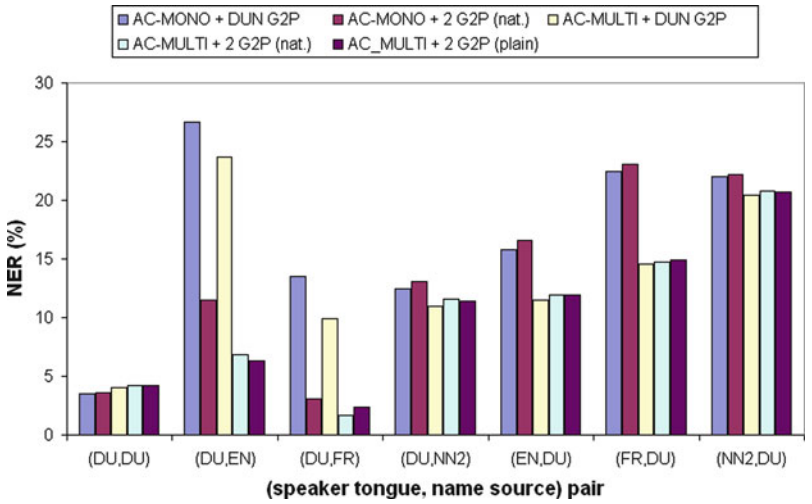

Fig. 14.1 NER results per ASNC cell for five different systems which differ in (a) the acoustic model (monolingual = AC-MONO, multilingual = AC-MULTI), (b) the G2P transcriptions included in the lexicon (DUN G2P = only a Dutch transcription, 2 G2P = additional English/French transcription for English/French names), and (c) the use of plain or natvised foreign G2P transcriptions

was published for the first time in [18]. It suggests that native speakers articulate foreign sounds with a native accent.

Based on the above conclusions we defined a state-of-the-art baseline system against which we will measure the effect of our lexical modeling approaches. Our baseline comprises a multilingual acoustic model (AC-MULTI) and a lexicon of pronunciations emerging from a Dutch, a French and an English G2P converter, in which the foreign transcriptions are natvised.

14.3 Potential for Further Improvement

The former experiments tell us what can be achieved with a lexical model based on existing general-purpose G2P converters. But what would a more advanced model be able to achieve? Imagine for instance that the lexicon contains for each name all actually used transcriptions of that name. How good would the recognition be then?

To test this situation, we supplemented the baseline lexicon with all auditorily verified transcriptions that were found in the training and test utterances of the ASNC. This resulted in a lexicon with 8.7 transcriptions per name on average. The improvements obtained with this lexicon (Table 14.2) were substantial for all cells. This makes it plausible that lexical modeling is able to yield a significant improvement over the baseline system.

Table 14.2 NER (%), per name source and per speaker tongue, for the baseline system and for a system with a lexicon that also comprises all actually used pronunciations per name

System	(DU,DU)	(DU,EN)	(DU,FR)	(DU,NN2)	(DU,ALL)
Baseline	4.2	6.8	1.7	11.6	5.5
Cheat	2.8	2.8	1.4	1.5	2.5
System	(DU,DU)	(EN,DU)	(FR,DU)	(NN2,DU)	(ALL,DU)
Baseline	4.2	11.9	14.7	20.8	10.6
Cheat	2.8	3.4	6.4	9.6	4.9

14.4 A Novel Pronunciation Modeling Approach

The proposed method creates pronunciation variants on the basis of automatically derived stochastic transformation rules that convert the phonemic output of a standard G2P into transcriptions that are more appropriate for names. Each rule predicts with which probability a phoneme sequence (called the *focus*) appearing in the initial G2P transcription (called the *source* transcription) may be phonetised as an alternative phoneme sequence (called the *rule output*) when it occurs in a particular linguistic context that can be defined in a flexible way (see below). The rules for a certain focus are embedded in the leaf nodes of a decision tree that uses yes/no-questions to distinguish between different contexts. Since the rules are stochastic in nature they will lead to multiple transcriptions per name with different probabilities. Although the VoCon engine cannot cope with these probabilities in the recognition lexicon,[4] they are still used for pronunciation selection during the lexicon creation. The presented approach constitutes a unique combination of the following features:

1. The transformable objects can be phonemic sequences (phoneme patterns) of different lengths (most published methods are confined to single phonemes).
2. The linguistic context is not restricted to the phonemic context (as in many other studies) but it can also include orthographic (graphemic), syllabic, morphological, syntactic and semantic information in a flexible way.
3. The computer-aided identification of suitable syllabic and morphological features is facilitated by built-in automatic procedures in the rule learning process.
4. The relevant (focus, output) combinations as well as the rules are learned fully automatically.

Other published methods (e.g. [12, 19–21]) share some of the above features, but we believe to be the first to propose and assess a method incorporating all these features simultaneously.

[4]This is an unfortunate limitation of the VoCon recogniser. Estimates based on preliminary experiments in which VoCon N-best hypothesis lists were rescored with the transcription variant probabilities learn that the latter can probably bring additional gains of up to 5 % relative.

The derived rules constitute a so-called P2P converter. It can be learned with the tools that were created in the first Autonomata project. All that is needed is a lexical database comprising of the order of a thousand names representative of the envisaged application domain. Per name, this database has to supply one or more plausible pronunciations and, optionally, some semantic tags (e.g. the name category). We argue that in many practical situations, such a database can be created cheaply because of its limited size, and because it can be elicited from one or two persons who are acquainted with the domain (and are able to write phonetics). These persons can select the names and enter their typical pronunciations.

Since the typical transcriptions have to be supplied by a human, the method as a whole is only semi-automatic, but once all transcriptions are available, the method is conceptually automatic. Nevertheless, it is practically implemented as a process that permits the user to intervene in an easy and transparent way if he believes that with these interventions he can surpass the improvements attainable with the automatic procedure. Note that the interventions boil down to simple updates of text files on the basis of statistical information that is being generated automatically after each step of the rule learning procedure.

Let us now review the different steps of our method, starting with a review of the contextual features we have selected.

14.4.1 Contextual Features

First of all, we consider the two phonemes immediately preceding and succeeding the focus as the primary contextual features (= 4 features). However, as in [19], we also take syllabic information into account, such as the identities of the vowels of the focus syllable and its two surrounding syllables (= 3 features) and the stress levels (no stress, primary stress or secondary stress) of these syllables (= 3 features).

Secondly, we follow the argument of Schaden [20, 21] that the orthography plays a crucial role in non-native pronunciation variation modeling because it is the key to the detection of systematic phonetisation errors. Take the French cheese name “Camembert” for instance. While the native pronunciation of this name is /“ka.ma˜.bER/, a native Dutch speaker may be inclined to pronounce it as /ka.m@m.“bErt/ because in Dutch, a “t” in the orthography is normally not deleted in the pronunciation (cf. [21] for more examples). The main limitation of Schaden’s work was that it employed handcrafted rules. In a similar vein, [12] incorporated graphemic information in an automatic data-driven approach, but the limitation of that work was that the focus had to be a single phoneme and that the graphemic context was restricted to the grapheme that gave rise to this focus. For our experiments, we considered four graphemic features: the graphemic pattern that caused the focus (but restricted to the first two graphemic units), the graphemic units immediately to the left and the right of this pattern, and a flag signaling whether or not the graphemic pattern causing the focus ends on a dot (= a simple indicator of an abbreviation).

Thirdly, we support the suggestion of Schaden [21] to consider morphological information as a potentially interesting context descriptor. Schaden noticed for instance that the vowels in the German suffixes "-stein" and "-bach" are less susceptible to accented pronunciations than the same vowels in other morphological contexts, but he did not actually build a system exploiting this observation. Since we would need multiple morphological analyzers in our cross-lingual setting, since these analyzers are expected to fail on many proper names and since we believe that a detailed morphological analysis is not very effective for our purposes, we did not try to incorporate them. Instead, we opted for a simple and pragmatic approach which automatically detects syllables, prefixes and suffixes often co-occurring with name transcription errors:

1. Three booleans indicating whether the focus syllable, the previous and the next syllable belong to a user-specified list of problematic syllables,
2. A boolean indicating whether the focus appears in a word starting with a prefix that belongs to a user-specified prefix list,
3. A boolean indicating whether the focus appears in a word ending on a suffix that belongs to a user-specified suffix list,
4. The positions (in numbers of syllables) of the focus start and end w.r.t. the first and last syllable of the name stem respectively (the name stem is obtained by depriving the name of the longest prefix and suffix from the user-specified prefix and suffix lists).[5]

Further below we will explain how to get the mentioned syllable, prefix and suffix lists in a semi-automatic way.

Finally, we believe that in the envisaged applications of proper name recognition, high-level semantic information such as the name category (e.g. street name, city name, Point of Interest), the source of the inquired name (if known), etc. are important to create more dedicated pronunciation variants. Therefore, we devised the P2P learning software so that such semantic tags can be accommodated through boolean features that are true if the tag belongs to predefined value sets (the values are character strings). In the experiments that will be discussed later, we employed the name category as a semantic feature, while the language of origin (which was supposed to be given) was used to select the proper P2P converter (e.g. the one intended for English names spoken by Dutch speakers).

14.4.2 The Overall Rule Induction Process

Since the phonemic focus patterns and the contextual features for the rule condition are not a priori known, the rule induction process is a little more complicated than usual. The process is outlined in Fig. 14.2. In general terms, the process is applied

[5] If the focus starts/ends in the selected prefix/suffix, the corresponding position is zero.

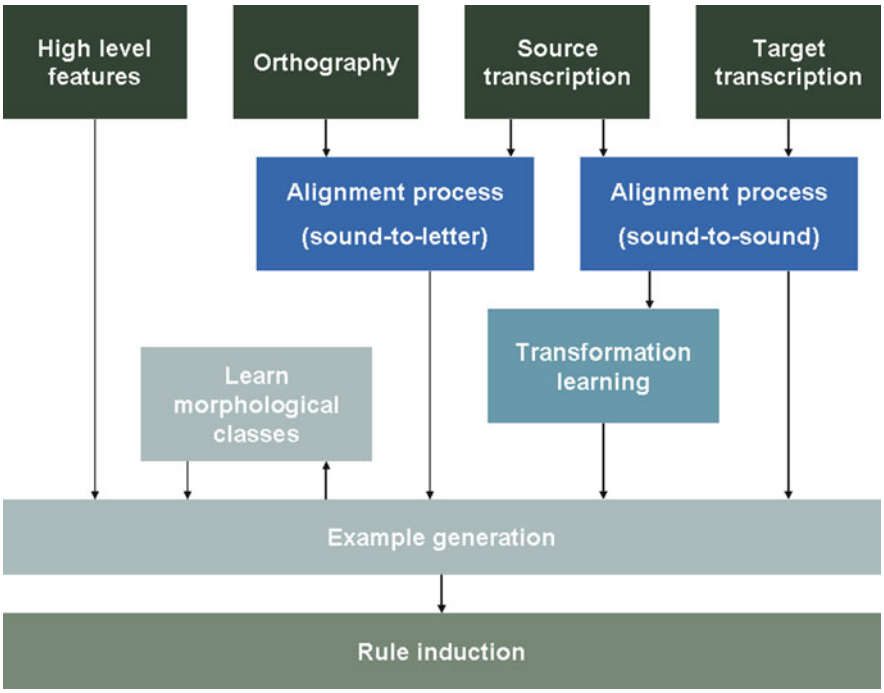

Fig. 14.2 Process for the automatic learning of a P2P converter

to a set of training objects each consisting of an orthography, a source transcription, a target transcription and a set of high-level features. Given these training objects, the learning process then proceeds as follows:

1. The objects are supplied to an alignment process incorporating two components: one for lining up the source transcription with the target transcription (sound-to-sound) and one for lining up the source transcription with the orthography (sound-to-letter). These alignments, together with the high-level features (morphological and semantic features) are stored in an alignment file.
2. The transformation learner analyzes the alignments and identifies the (focus, output) pairs that are capable of explaining a sufficiently large number of deviations between the source and the target transcriptions. These pairs are stored in a transformation file from which one can obviously retrieve the focus patterns.
3. The alignment file and the transformation file are supplied to the example generator. The latter searches for focus patterns in the source transcriptions and it generates a file containing the focus, the corresponding contextual features and the output for each detected focus pattern. These combinations will serve as the examples from which to train the rules. If no morphological features have been defined yet, one can define them on the basis of statistical information produced by the example generator. After that, one can run the example generator a second

time to create the final training examples that will also incorporate these features then.
4. The example file is finally supplied to the actual rule induction process that automatically constructs a binary decision tree per focus. Each tree is grown incrementally by choosing per leaf node the yes/no question leading to the largest entropy loss and by accepting the resulting split if this loss exceeds a predefined threshold. The rule probabilities can be derived from the counts of the different eligible outputs in each leaf node of the tree.

The full details of the approach are described in Chap. 4 of this book (cf. Sect. 4.3, p. 67) and in a journal paper [17]. We just mention here that the statistical information provided by the example generator reveals the number of co-occurrences of a discrepancy between the source and the target transcription and a syllable identity or a word property. The two word properties being considered are the graphemic sequences that correspond to the first and last one or two syllables of the word respectively. For instance, if a discrepancy frequently appears in a word starting with "vande", this "vande" will occur in the word prefix list.

14.5 Experimental Validation

In this section we investigate under which circumstances the proposed lexical methodology can enhance the name recognition performance. We first conduct experiments on the ASNC that covers the person and topographical name domains. Then, we verify whether our conclusions remain valid when we move to another domain, in casu, the POI domain.

14.5.1 Modes of Operation

Since in certain situations it is plausible to presume prior knowledge of the speaker tongue and/or the name source, three relevant modes of operation of the recogniser are considered:

- **M1**: In this mode, the speaker tongue and the source of the inquired name are a priori known. That is, the case of a tourist who uses a voice-driven GPS system to find his way in a foreign country where the names (geographical names, POI names) all originate from the language spoken in that country.
- **M2**: In this mode, the speaker tongue is known but names from different sources can be inquired. Think of the same tourist who is now traveling in a multilingual country like Belgium where the names can either be Dutch, English, French, German, or a mixture of those.

Table 14.3 NER (%), per name source and per speaker tongue, obtained with multilingual acoustic models and three distinct lexicons: (a) the baseline lexicon (2 G2P), (b) a lexicon also comprising variants generated by a P2P converter trained on the ASNC training names (ASNC), and (c) a lexicon also comprising variants generated by a P2P converter trained on an extended name set (ASNC+).

System	(DU,DU)	(DU,EN)	(DU,FR)	(DU,NN2)	(DU,ALL)
baseline (AC-MULTI + 2 G2P-nat)	4.2	6.8	1.7	11.6	5.5
baseline + 4 P2P variants (ASNC)	3.8	5.3	1.7	6.2	4.2
baseline + 4 P2P variants (ASNC+)	–	4.7	1.4	–	4.1
System	(DU,DU)	(EN,DU)	(FR,DU)	(NN2,DU)	(ALL,DU)
baseline (AC-MULTI + 2 G2P-nat)	4.2	11.9	14.7	20.8	10.6
baseline + 4 P2P variants (ASNC)	3.8	10.2	12.5	19.5	9.6

- **M3**: In this mode, neither the mother tongue of the actual user nor the source of the inquired name are a priori known. This mode applies for instance to an automatic call routing service of an international company.

The first experiments are carried out under the assumption of mode M1. In that case, we know in which cell we are and we only add variants for names that can occur in that cell. Furthermore, we can in principle use a different P2P converter in each cell. However, since for the ASNC names we only had typical native Dutch transcriptions, we could actually train only four P2P converters, one per name source. Each P2P converter is learned on a lexical database containing one entry (orthography + Dutch G2P transcription + typical Dutch transcription) per name of the targeted name source.

14.5.2 Effectiveness of P2P Variants

After having evaluated the transcription accuracy improvement as a function of the number of selected P2P variants, we came to the conclusion (cf. [17]) that it is a viable option to add only the four most likely P2P variants to the baseline lexicon. By doing so, we obtained the NERs listed in Table 14.3.

The most substantial improvement (47 % relative) is obtained for the case of Dutch speakers reading NN2 names. For the case of Dutch speakers reading French names no improvement is observed. The gains in all other cells are more modest (10–25 % relative), but nevertheless statistically significant ($p < 0.05$, even $p < 0.01$ for Dutch and NN2 names uttered by Dutch speakers.[6])

The fact that there is no gain for native speakers reading French names is partly owed to the fact that the margin for improvement was very small (the baseline 2

[6]Statistical significance of NER differences is determined using the Wilcoxon signed ranks test [22].

G2P system only makes seven errors in that cell, cf. also Table 14.2). Furthermore, the number of examples that is available for the P2P training is limited for French names. While there are 1,676 training instances for Dutch names, there are only 322 for English names, 161 for French names and 371 for NN2 names. Therefore, we performed an additional experiment in which the sets of English and French training names were extended with 684 English and 731 French names not appearing in the ASNC test set. The name set including these extensions is called ASNC+. Training on this set does lead to a performance gain for French names. Moreover, the gain for English names becomes significant at the level of $p < 0.01$ (cf. Table 14.3).

In summary, given enough typical transcriptions to train a P2P converter, our methodology yields a statistically significant ($p < 0.01$) reduction of the NER for (almost) all cells involving Dutch natives. For the utterances of non-natives the improvements are only significant at the level of $p < 0.05$ for speakers whose mother tongue is covered by the acoustic model. This is not surprising, since the Dutch typical transcriptions that we used for the P2P training were not expected to represent non-native pronunciations. Larger gains are anticipated with dedicated typical training transcriptions for these cells.

14.5.3 Analysis of Recognition Improvements

Our first hypothesis concerning the good results for native speakers was that for these speakers, there is not that much variation to model within a cell. Hence, one single TY transcription target per name might be sufficient to learn good P2P converters. To verify this hypothesis we measured, per cell, the fraction of training utterances for which the auditorily verified transcription is not included in the baseline G2P lexicon. This was the case for 33 % of the utterances in cell (DU,DU), around 50 % in (DU,EN) and (DU,FR) and around 75 % in all other cells, including (DU,NN2) for which we also observed a big improvement.

The small improvement achieved for NN2 speakers reading Dutch names is owed to the fact that many NN2 speakers have a low proficiency in Dutch reading, which implies that they often produce very a-typical phonetisations. The latter are not modeled by the Dutch typical transcriptions in our lexical database. Another observation is that NN2 speakers often hesitate a lot while uttering a native name (cf. [23]) and these hesitations are not at all modeled either.

In order to find an explanation for the good results for Dutch speakers reading NN2 names, we have compared two sets of P2P converters: one trained towards typical transcriptions and one trained towards ideal (auditorily verified) transcriptions as targets. We have recorded how many times the two P2P converters correct

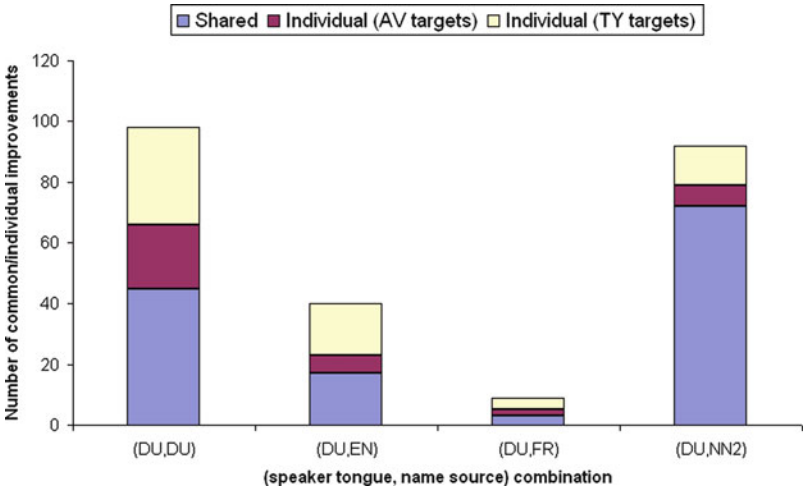

Fig. 14.3 Number of error corrections that can be achieved with the variants generated by the P2P converter trained towards typical (TY) targets, the P2P converter that was trained towards auditorily verified (AV) targets, and both P2P converters (shared)

the same recognition error in a cell and how many times only one of them does. Figure 14.3 shows the results for the four cells comprising Dutch speakers.[7]

It is remarkable that in cell (DU,NN2) the percentage of errors being corrected by both P2P converters is significantly larger than in the other cells. Digging deeper, we came to the conclusion that most of these common corrections were caused by the presence of a small number of simple vowel substitution rules that are picked up by both P2P converters as they represent really systematic discrepancies between the G2P and the typical transcriptions. The most decisive rules express that the frequently occurring letter "u" in NN2 names (e.g. Curukluk Sokagi, Butrus Benhida, Oglumus Rasuli, etc.) is often pronounced as /u/ (as in "b**oo**t") while it is transcribed as /Y/ (like in "m**u**d") or /y/ (like in the French "cr**u**") by the G2P converter.

Similarly, we have also examined for which names the P2P variants make a positive difference in the other cells. Table 14.4 gives some representative examples of names that were more often correctly recognised after we added P2P variants.

An interesting finding (Table 14.4) is that a minor change in the name transcription (one or two phoneme modifications) can make a huge difference in the recognition accuracy. The insertion of an /n/ in the pronunciation of "Duivenstraat" for instance leads to five corrected errors out of six occurrences.

[7]Note that we actually obtained these results with a system comprising a larger recognition vocabulary of 21K person and geographical names. For more details we refer to [17].

Table 14.4 Examples of proper names for which the recognition improves. Listed are: (a) the name, (b) its baseline transcription(s), (c) the P2P variant that led to an error reduction, (d) the netto number of improvements versus the number of occurrences of a name

Name	Baseline G2P variant(s)	Helping P2P variant	Netto positive result
Duivenstraat	"d9y.v@.stra:t	"d9y.v@n.stra:t	5/6
Berendrecht	b@."rEn.drExt	"be:.rEn.drExt	4/6
Carter Lane	"kAr.t@r#"la:.n@ "kA.t@#"le:jn	"kAr.t@r#"le:.n	3/6
Norfolk	nOr."fOlk "nO.f@k	"nOr.fOk	3/6
Middlesbrough	"mIt.l@z.brux "mI.d@lz.br@	"mI.d@lz.bro:	2/6
Engreux	EN."r2:ks a˜."gr2:	EN."r2:	2/6
Renée Bastin	r@."ne:#bAs."tIn r@."ne:#ba:s."te˜	rE."ne:#bAs."te˜	3/6

Table 14.5 NER results (%) for names of different sources spoken by Dutch speakers. Shown are the results for the baseline system, the best P2P system under mode M1 and the results of the P2P system under mode M2

System	(DU,DU)	(DU,EN)	(DU,FR)	(DU,NN2)	(DU,ALL)
2 G2P, mode M1	4.2	6.8	1.7	11.6	5.5
2 G2P + 4 P2P, mode M1	3.8	4.7	1.4	6.2	4.1
2 G2P + 4 P2P, mode M2	4.0	4.9	2.2	6.9	4.4

14.5.4 Effectiveness of Variants in Mode M2

So far, it was assumed that the recogniser has knowledge of the mother tongue of the user and the origin of the name that will be uttered (mode M1). In many applications, including the envisaged POI business service, a speaker of the targeted group (e.g. the Dutch speakers) can inquire for names of different origins. In that case, we can let the same P2P converters as before generate variants for the names they are designed for, and incorporate all these variants simultaneously in the lexicon. With such a lexicon we got the results listed in Table 14.5. For the pure native situation, the gain attainable under mode M2 is only 50 % of the gain that was achieved under mode M1. However, for cross-lingual cases (apart from the French names case), most of the gain achieved under mode M1 is preserved under mode M2. Note that in case of the French names, the sample size is small and the difference between 1.4 and 2.2 % is only a difference of three errors.

Table 14.6 NER results (%) for Dutch name spoken by non-native speakers. Shown are the results for the baseline system, the best P2P system under mode M1 and the results of the P2P system under mode M2

System	(DU,DU)	(EN,DU)	(FR,DU)	(NN2,DU)	(ALL,DU)
2 G2P, mode M1	4.2	11.9	14.7	20.8	10.6
2 G2P + 4 P2P, mode M1	3.8	10.2	12.5	19.5	9.6
2 G2P + 4 P2P, mode M2	4.0	10.9	13.9	20.2	10.1

14.5.5 Effectiveness of Variants in Mode M3

In case neither the mother tongue of the speaker nor the origin of the name is given beforehand (mode M3), the recognition task becomes even more challenging. Then variants for all name sources and the most relevant speaker tongues have to be added at once.

Since we had no typical non-native pronunciations of Dutch names at our disposal, a fully realistic evaluation of mode M3 was not possible. Consequently, our lexicon remained the same as that used for mode M2, meaning that the results for native speakers remain unaffected. The results for the non-native speakers are listed in Table 14.6. They are put in opposition to the baseline results and the results with lexical modeling under mode M1. The figures show that in every cell about 50 % of the gain is preserved. This implies that lexical modeling for proper name recognition in general is worthwhile to consider.

14.5.6 Evaluation of the Method in the POI Domain

In a final evaluation it was verified whether the insights acquired with person and typographical names transfer to the new domain of POI. For the training of the P2P converters we had 3,832 unique Dutch, 425 unique English and 216 unique French POI names available, each delivered with one or more plausible native Dutch transcriptions[8] and a language tag. Since there was a lot less training material for French and English names, we also compiled an extended dataset (POI+) by adding the French and English training instances of the ASNC+ dataset.

For the experimental evaluation of our method, we used the POI name corpus that was created in Autonomata Too, and that is described in Chap. 4 of this book (cf. Sect. 4.5, p. 74) and in [23].

Here we just recall that Dutch speakers were asked to read Dutch, English, French and mixed origin (Dutch-English, Dutch-French) POI, while foreign speakers were asked to read Dutch and mixed origin POI only. The recordings were

[8]The number of actual training instances per language was 6,681 for Dutch, 991 for English and 486 for French.

Table 14.7 NER results (%) for POI of different sources spoken by Dutch speakers. Shown are the results for the baseline system (2 G2P) and the P2P systems under modes M1 and M2

System	(DU,DU)	(DU,EN)	(DU,FR)	(DU,ALL)
2 G2P, mode M1	7.7	7.8	9.6	8.5
2 G2P + 4 P2P, mode M1 (POI)	6.6	6.9	8.4	7.5
2 G2P + 4 P2P, mode M1 (POI+)	–	6.9	8.1	7.3
2 G2P + 4 P2P, mode M2 (POI)	6.8	7.7	9.3	8.2
2 G2P + 4 P2P, mode M2 (POI+)	–	7.6	9.0	8.1

Table 14.8 NER results (%) for Dutch POI spoken by non-native speakers. Shown are the results for the baseline system (2 G2P) and the P2P systems under modes M1 and M2

System	(DU,DU)	(EN,DU)	(FR,DU)	(NN2,DU)	(ALL,DU)
2 G2P, mode M1	7.7	13.6	8.8	22.8	15.0
2 G2P + 4 P2P, mode M1	6.6	13.0	8.4	22.1	14.3
2 G2P + 4 P2P, mode M2	6.8	13.0	8.4	22.6	14.5

conducted such that the emphasis was on the cases of Dutch natives reading foreign POI and on non-natives reading Dutch POI.

The vocabulary of the recogniser consisted of 10K POI: all POI spoken in the POI name corpus, supplemented with additional POI that were drawn from background POI lexica provided by TeleAtlas. There was no overlap between this vocabulary and the POI set that was available for P2P training. Also, none of the POI occur in the ASNC.

Table 14.7 shows NER results for Dutch utterances under the assumptions of modes M1 and M2 respectively. Table 14.8 depicts similar results for the non-native speakers.

The data support the portability of our methodology. Adding P2P variants for POI in mode M1 strongly reduces the NER for Dutch native speakers and modestly improves the recognition for non-native speakers. In mode M2, the over-all result still holds that a substantial part of the gain is preserved. However, there are differences in the details. We now see a good preservation of the gain obtained in the purely native case, but the gains in the cross-lingual settings are more diverse. The preserved gain ranges from only 22 % (for Dutch speakers reading English names, with an extended training set) to 100 % (for English and French speakers reading Dutch names).

Furthermore, we see how an extended training set for English and French POI yields no improvement for English POI and only a small gain for French POI. This either reflects that the ASNC proper name transcriptions are not suited as training material for POI names, or that relevant information regarding the "correct" transcription of proper names can already be captured with a limited training set of name transcriptions. To verify the latter hypothesis, we performed two additional mode M1 recognition experiments for Dutch POI in which only one fourth (corresponding to about 1K unique names, 1.7K training instances) and one sixth (corresponding to about 1K training instances, for nearly 650 unique

training names) of the training set names for Dutch POI were included for the P2P converter training. We found that for both set-ups the NER was even (slightly) lower than before (6.4 % for 1K unique training POI names and 6.5 % for 1K training instances). We therefore argue that a limited training set of around 1K transcribed training names will typically be sufficient to learn a good P2P converter.

A qualitative evaluation of the improvements induced by the P2P transcriptions has been performed as well and is described in [24]. That evaluation confirmed that relatively simple phoneme conversions (substitutions, deletions, insertions) account for most of the obtained NER gains, but that a large number of more structural variations (e.g. syllable-size segment deletion) is not modeled by the P2P converters. An explicit modeling of these variations, possibly by means of other techniques, could further raise the efficiency of the POI recogniser.

14.6 Conclusions

We have proposed a novel lexical modeling methodology for the automatic recognition of proper names in a monolingual and cross-lingual setting. The method was experimentally assessed and compared against a baseline incorporating existing acoustic and lexical modeling strategies that have been applied to the same problem.

Our assessment of existing methodologies demonstrated that in a cross-lingual setting, proper name recognition can benefit a lot from a multilingual acoustic model and from transcriptions emerging from foreign G2P transcribers. We have further established that the two strategies are complementary.

The newly presented lexical modeling approach is unique in its combination of interesting properties that have never been integrated in a single system. Some of these features are: the transformation of variable length phonemic patterns from a baseline transcription, the extensive use of linguistic context at multiple levels (from phonemic to semantic), the computer-assisted identification of syllabic and morphological features, the automatic learning of context-dependent stochastic rules embedded in multiple decision trees, etc. An important feature of the method is that it does not need any labeled speech data as training material nor any expertise in automatic speech recognition. The downside is of course that the user must provide a lexical database of correspondences between a name and its typical transcription. However, since the required database is small (of the order of a thousand names), it is easy and cheap to construct.

The new method was evaluated under different modes of operation differing in the a priori knowledge one has about the mother tongue of the speaker and the language of origin of the name the speaker can inquire. When both languages are a priori known, one can achieve important reductions of the name error rate: from 10 % relative for the pure native setting, over 15 % relative for the cross-lingual settings involving a non-native language that was involved in the construction of the baseline lexicon and in the training of the multilingual acoustic models, to 45 % relative for the case where Dutch speakers read non-native names of a language

they are not familiar with. Note that the proposed method is currently not able to cope with the hesitations and strongly a-typical pronunciations of Dutch names by speakers with a low proficiency in Dutch.

References

1. Cremelie, N., ten Bosch, L.: Improving the recognition of foreign names and non-native speech by combining multiple grapheme-to-phoneme converters. In: Proceedings ISCA ITRW on Adaptation Methods for Speech Recognition, Sophia Antopolis, France, pp. 151–154 (2001)
2. Eklund, R., Lindstrom, R.: How foreign are 'foreign' speech sounds? Implication for speech recognition and speech synthesis. In: Proceedings RTO Meeting on Multi-Lingual Interoperability in Speech Technology, Hull, Canada, pp. 15–19 (1999)
3. Flege, J.: The production and perception of foreign language speech sounds. In: Winitz, H. (ed.) Human Communication and Its Disorders. A Review, pp. 224–401. Norwood, Ablex (1988)
4. Schaden, S.: Regelbasierte Modellierung fremdsprachlich akzentbehafteter Aussprachevarianten. PhD Dissertation University of Bochum (2006)
5. Trancoso, I., Viana, C., Mascarenhas, I., Teixeira, C.: On deriving rules for nativised pronunciation in navigation queries. In: Proceedings Eurospeech, Budapest, Hungary, pp. 195–198 (1999)
6. Maison, B., Chen, S., Cohen, P.: Pronunciation modeling for names of foreign origin. In: Proceedings ASRU, Virgin Islands, USA, pp. 429–434 (2003)
7. Schultz, T., Kirchhof, K.: Multilingual Speech Processing. Elsevier, Academic (2006)
8. Stouten, F., Martens, J.: Recognition of foreign names spoken by native speakers. In: Proceedings Interspeech, Antwerp, Belgium, pp. 2133–2136 (2007)
9. Leggetter, C., Woodland, P.: Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. Comput. Speech Lang **9**, 171–185 (1995)
10. Gauvain, J.-L., Lee, C.-H.: Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. IEEE Trans. Speech Audio Process. **2**, 291–298 (1994)
11. Mayfield-Tomokiyo, L., Waibel, A.: Adaptation methods for non-native speech. In: Proceedings Workshop on multilinguality in Spoken Language Processing, Aalborg, Denmark (2001)
12. Bouselmi, G., Fohr, D., Illina, I., Haton, J.: Fully automated non-native speech recognition using confusion-based acoustic model integration and graphemic constraints. In: Proceedings ICASSP, Toulouse, France, pp. 345–348 (2006)
13. Stemmer, G., Nöth, E., Niemann, H.: Acoustic modeling of foreign words in a german speech recognition system. In: Proceedings Eurospeech, Aalborg, Denmark, pp. 2745–2748 (2001)
14. Li, Y., Fung, P., Xu, P., Liu, Y.: Asymmetric acoustic modeling of mixed language speech. In: Proceedings ICASSP, Prague, Czech Republic, pp. 5004–5007 (2011)
15. Bartkova, K., Jouvet, D.: Using multilingual units for improving modeling of pronunciation variants. In: Proceedings ICASSP, Toulouse, France, pp. 1037–1040 (2006)
16. Bartkova, K., Jouvet, D.: On using units trained on foreign data for improved multiple accent speech recognition. Speech Commun. **49**(10–11), 836–846 (2007)
17. Reveil, B., Martens, J., van den Heuvel, H.: Improving proper name recognition by means of automatically learned pronunciation variants. Speech Commun. **54**(3), 321–340 (2012)
18. Réveil, B., Martens, J.-P., D'Hoore, B.: How speaker tongue and name source language affect the automatic recognition of spoken names. In: Proceedings Interspeech, Brighton, UK, pp. 2995–2998 (2009)

19. Riley, M., Byrne, W., Finke, M., Khudanpur, S., Ljolje, A., McDonough, J., Nock, H., Saraclar, M., Wooters, C., Zavaliagkos, G.: Stochastic pronunciation modelling from hand-labelled phonetic corpora. Speech Commun. **29**(2–4):209–224 (1999)
20. Schaden, S.: Rule-based lexical modelling of foreign-accented pronunciation variants. In: Proceedings 10th EACL Conference, Budapest, Hungary, pp. 159–162 (2003)
21. Schaden, S.: Generating non-native pronunciation lexicons by phonological rules. In: Proceedings ICPhS, Barcelona, Spain, pp. 2545–2548 (2003)
22. Conover, W.: Practical Nonparametric Statistics, vol. 3. Wiley, New York (1999)
23. Schraagen, M., Bloothooft, G.: Evaluating repetitions, or how to improve your multilingual asr system by doing nothing. In: Proceedings LREC, Valletta, Malta, pp. 612–617 (2010)
24. Schraagen, M., Bloothooft, G.: A qualitative evaluation of phoneme-to-phoneme technology. In: Proceedings Interspeech, Florence, Italy, pp. 2321–2324 (2011)