G2P conversion of names. What can we do (better)?

Henk van den Heuvel[1], *Jean-Pierre Martens*[2], *Nanneke Konings* [1]

[1] CLST, Radboud University Nijmegen, The Netherlands
[2] ELIS, Ghent University, Belgium

H.vandenHeuvel@let.ru.nl

Abstract

In this contribution it is shown that a good approach for the grapheme-to-phoneme conversion of proper names (e.g. person names, toponyms, etc), is to use a cascade of a general purpose grapheme-to-phoneme (G2P) converter and a special purpose phoneme-to-phoneme (P2P) converter. The G2P produces an initial transcription that is then transformed by the P2P. The latter is automatically trained on reference transcriptions of names belonging to the envisaged name category (e.g. toponyms). The P2P learning process is conceived in such a way that it can take account of high order determinants of pronunciation, such as specific syllables, name prefixes and name suffixes. The proposed methodology was successfully tested on person names and toponyms, but we believe that it will also offer substantial reductions of the cost for building pronunciation lexicons of other name categories.

Index Terms: G2P conversion, names, Dutch, machine based learning

1. Introduction

Correct phonetic transcriptions are of paramount importance for both automatic speech-to-text (STT) conversion and text-to-speech (TTS) conversion. The manual generation of these transcriptions is very time-consuming and subject to a great deal of inconsistencies. For this reason, automatic grapheme-to-phoneme (G2P) converters have been developed [1, 2, 3, 6, 9, 11]. However, general purpose converters often perform poorly when it comes to the transcription of names. Names typically do not adhere to the standard spelling conventions of a language due to their fossilized orthographic forms and/or their foreign origin. Therefore, they need special treatment [3, 7, 12].

In the AUTONOMATA[1] project we have chosen for an approach in which an initial transcription, emerging from a state-of-the-art general-purpose G2P converter, is 'corrected' by a special-purpose phoneme-to-phoneme (P2P) converter. This is an attractive option because it permits the P2P converter to profit from the knowledge of the general-purpose G2P converter, and it can focus on pronunciation rules that are typical for the envisaged name category. As such, it can be compact (few rules), and trainable on a relatively small pronunciation dictionary comprising only a few thousand names with their manually verified transcriptions.

Autonomata has delivered G2P-P2P tandems for person names and toponyms, as well as a *methodology* for the creation of good P2P's for other word classes for which a standard G2P is known to perform poorly (e.g. brand names, Points-Of-Interests, etc.). We have explored two different approaches. The first one is called the *inductive* approach. An inductive rule learning algorithm retrieves stochastic rules (attached to the leaf nodes of decision trees) that can explain many of the discrepancies between the correct transcription and the transcription produced by the standard G2P (in terms of phonemes and their immediate contexts). In the second approach, human expert knowledge is used in a top-down fashion to relate errors to higher order (viz. morphology, affix splitting, and language origin). For example, in street names we can identify morphemic entities such as 'erf' (yard), 'kamp' (parcel), 'straat' (street) which must be taken into account for the production of a proper syllabification, stress assignment and phonemization. We qualify this second approach as *deductive*.

In this study we present two series of experiments that were conducted in order to compare the inductive and deductive approaches. For the first experimental series (ES1) we asked ourselves the following questions:

1. What are the performances of the deductive and inductive approaches when used independently?
2. What is the range of improvements that the synergy of both approaches can bring when compared to the inductive approach alone?
3. What are the most relevant features the deductive approach can add to the inductive approach?

In the second series of experiments (ES2) we focused on the potential synergy of both approaches. The main research questions were:

1. Is it possible to incorporate the findings of the deductive approach into the P2P learning software?
2. Are additional improvements achievable by extracting and adding further deductive rules?
3. What is the effect of the training set size?
4. Can the G2P-P2P cascade perform as well as a dedicated G2P that was trained to generate a phoneme transcription of the names directly from the orthography?

We will only briefly report on the results of ES1 to leave more space for presenting ES2 in more detail.

2. Materials and tools

The reference lexicons that were available in the Autonomata project included first names, last names and toponyms (place names and, mostly, street names) that are encountered in the Netherlands and in Flanders. In the present study however, only names from the Netherlands were considered. Each lexicon was divided in a training, a development and an evaluation set [10]. Table 1 shows the sizes of these sets for the various name types.

Table 1. *Sizes of the manually verified reference lexicons*

Lexicon	Train set	Dev. set	Eval. set
First names	15,655	3,913	2,987
Last names	64,048	16,011	8,382
Toponyms	92,838	23,209	12,483

For each name, one or more manual phoneme transcriptions were available as reference transcriptions during training and evaluation. The manual transcriptions are broad phonetic transcriptions, enriched by syllable and stress markers.

The general-purpose G2P was provided by Nuance. The inductive learning toolbox was developed by ELIS and is described in [12]. However, some important extensions were implemented since then. They are described in Section 4.

For the deductive approach we used internal tools of CLST. Based on different contexts of one, two or three phonemes to the left and/or right of the target phonemes, we first identified (in the training set) the most frequent erroneous input patterns and their corresponding outputs, and we computed rule application rates for these input-output pairs. The application rate is 1 if the input always corresponds to the same output. Next, we tried to relate the observed errors to higher order phenomena, more specifically to the morphological composition of the name. Correction rules were then formulated in the FONPARS format [8] and applied to the output of the G2P.

To evaluate the performance of the inductive and deductive rules (or the combination thereof) the G2P-P2P output transcription is aligned with the reference transcription and the number of mismatches is counted at the phoneme level and the word (=name) level. Based on this comparison the evaluation yields a symbol error rate (SER) and a word error rate (WER). The WER is the percentage of names with one or more errors in their transcriptions. The evaluation tool also counts the mismatches between the original G2P transcription and the reference transcription. This way we can compute the word improvement rate (WIR) as the percentage of words for which the P2P offers an improvement minus the percentage of words for which it causes a degradation of the transcription. A transcription is called better if it has less symbols that deviate from the reference transcription (even if it is still wrong, and thus leaving the WER unchanged). In our experiments we will report SER, WER, and WIR. Further, we will report on the number of rules that were needed to arrive at the result, the idea being that given the same accuracy, a small rule set is preferred over a large one.

Although language origin may be an important factor, this information is not presumed to be available. In fact, the positive effect of distinguishing the language origin is not so straightforward as it might seem at first glance. E.g. the name Johnny, which is clearly of an English origin, has an accepted 'dutchified' pronunciation /ʃɔni/.

Although the G2P-P2P cascades can be instructed to produce multiple transcriptions, the experiments reported here considered only *one* transcription per name. This transcription is then compared to the best matching one of the manual pronunciations available in the evaluation lexicon. Person names and toponyms were treated separately in the experiments.

3. Experiments ES1

As stated in the introduction we just briefly discuss the results for ES1. In the experiments we found that, for toponyms,

1. Both the inductive and deductive approach yield substantial improvements over the general-purpose G2P. The inductive approach performed better than the deductive approach (WIR of 27.4% vs. 20.6%).
2. Cascading deductive rules after the inductive P2P did yield a further improvement. This synergetic approach raised the WIR from 27.4% to 31.3% and lowered WER from 38.1% to 32.5%.
3. The most relevant findings of the deductive approach were that it helps to take the identity of syllables, prefixes and suffixes into account for computing a transcription. For toponyms, these syllables, prefixes and suffixes were:
   - **syllables**: *be de ge he ber der het ver kor laan sint sint-straat toor van*
   - **suffixes**: *de el ne se ter baan dijk dreef hof jitte laan meester pad plein singel straat weg destraat sestraat seweg steenweg*
   - **prefixes***: be de ge he ber ver bo ca ha ho ka ma mo ou pa ro burge sint-*

   For person names, the syllables, prefixes and suffixes were:
   - **syllables**: *de den der het te ten ter ver van a van*
   - **suffixes**: *je ske sje ke kje the na ne nus se ta us berg burg re ren sen de den en er ga gen ijer ker len man mans meijer ter veld ven zen*
   - **prefixes***: be ca ge ka ma de van vande vander ber ger*

4. Experiments ES2

In ES2 we investigated the synergetic potential of the deductive and inductive approach..

4.1. Extension of the inductive learning software

The original inductive learning software was extended so that it can take into account the phenomena discovered by the deductive approach in ES1. In the linguistic context of a pattern that is considered for modification (the rule focus), one can now include the syllable identity as well as the identities of the name prefix and suffix. These identities are either a syllable, prefix or suffix belonging to a predefined set, or *unknown*. The discovery of appropriate syllable, prefix and suffix sets is also automated. To that end, the learning software records all syllables, prefixes and suffixes it encounters, and it records how many times such an item occurs and how many times it co-occurs with a transcription error: a phonemization error in the syllable or an arbitrary error in a name with the given prefix or suffix. By just retaining the items with a sufficiently high co-occurrence rate, one can construct the syllable, prefix and suffix sets to supply to the learning software as part of the data file that also specifies the other linguistic features. One can even change the sets without having to change anything in the software. In this way, it is straightforward to apply the methodology on other name categories requiring other sets of prefixes, suffixes and syllables. In order to constrain the computation time and memory, the search for suitable prefixes and suffixes during the training stage is limited to syllables and syllable pairs as they are output by the general-purpose G2P

4.2. Experimental setup

To prove our case, we have compared our G2P-P2P approach against a direct approach in which a special purpose G2P, is trained on the same data the P2P was trained on. We have used TiMBL (Tilburg Machine Based Learning) [5] to train such G2P's. For TiMBL the number of rules is simply the number of training instances (=names). In all tests, we developed two systems per approach: one for converting person names and one for converting toponyms. During evaluation we separately tested on first names and last names. This leads us to the following experimental conditions:

1. **G2P** proper: baseline general-purpose G2P
2. **TiMBL** (trained on full train set): special purpose G2P trained on the full training set

3. **G2P + P2P(Ind):** inductive rules only, but now with syllable, prefix and suffix features included. To measure the effect of the training set size, we distinguished between systems trained on the full training set (**A**), and on the smaller development set (**B**).
4. **G2P + P2P(Ind) + P2P(Ded):** a combination of the previous system B and extra deductive rules obtained after having inspected (in the development set) the remaining errors made by this system.

The same experimental conditions were applied to both person names and toponyms for Dutch.

4.3. Results

The results of our experiments are presented in Table 2. The 95% confidence intervals of the WER scores were added to show the statistical significance of the differences. We can summarize our findings as follows:

- For first names none of the approaches leads to significant improvements, whereas for the two other categories all approaches result in significant improvements across all three measures (SER, WER and WIR).
- For toponyms for instance, one observes an extra gain of 3.5% absolute in the WIR, when compared with ES1.
- For all name categories, only small insignificant gains in performance are obtained by adding further deductive rules.
- The special purpose G2P's, developed with TIMBL, can only compete (in terms of accuracy) with the G2P-P2P systems for the case of family names.
- Comparing systems 3A and 3B for all three name types shows that the development set suffices for the training of a nearly optimal inductive P2P.

The results for family names are much better than for first names, probably because the training set for person names consisted of 75% last names and only 25% first names. Also the length of first names is considerably shorter, giving the P2P learning tool less context to fine-tune its adjustments.

5. Discussion & conclusions

Returning to our initial research questions for ES2 as formulated in the Introduction, we can draw the following conclusions:

1. Taking syllabic context and morphological context (affixes) discovered by a deductive approach into account during inductive learning, leads to an improved performance (as compared to the WIRs obtained in ES1 (section 3);
2. The obtained improvements must be close-to-optimal since the addition of more deductive elements does not yield any significant improvement anymore.
3. A relatively small development set suffices to train a nearly optimal P2P.
4. The comparison with TiMBL shows that the G2P-P2P approach is far more effective in terms of required training data, and transcription accuracy than a restart from scratch approach in which a special purpose G2P is trained from scratch on the same training data.

Except for first names, the G2P-P2P tandem yields substantial improvements over the G2P alone. Nonetheless, more than 30% of the name transcriptions still contain one or more errors. This makes a manual post-hoc correction necessary, be it less time consuming than before.
If the stress marks are excluded from the evaluation, the WER scores of the best G2P-P2P systems are about 30% for the

Table 2: *Results for ES2, for transcriptions containing both segmental and supra-segmental (syllabic, morphological) information. Results for Dutch names only. Number of extra rules refers to the extra rules per component on top of the G2P.*

Name type	System	# extra rules	SER	WER	WIR	95% CI on WER
First names	**1. G2P**	N/A	11.9	39.9	N/A	[38.1 – 41.7]
	2. TiMBL (Full training set)	(79.711)	12.4	52.5	-10.5	
	3A. G2P + P2P(Ind) (Full training set)	2434	10.5	40.4	5.4	
	3B. G2P + P2P(Ind) (development set)	2064	10.9	41.6	3.5	[39.8 – 43.4]
	4. G2P + P2P(Ind)+P2P(Ded) (development set)	**2064 + 28**	**10.7**	**40.9**	**4.3**	
Family names	**1. G2P**	N/A	9.5	44.6	N/A	[43.5 – 45.7]
	2. TiMBL (Full training set)	(79.711)	5.7	32.4	17.2	
	3A. G2P + P2P(Ind) (Full training set)	2434	6.5	35.5	20.3	
	3B. G2P + P2P(Ind) (development set)	2064	6.5	35.2	20.3	[34.2 – 36.2]
	4. G2P + P2P(Ind)+P2P(Ded) (development set)	**2064 + 28**	**6.3**	**34.4**	**21.3**	
Toponyms	**1. G2P**	N/A	6.8	51.2	N/A	[50.3 – 52.1]
	2. TiMBL (Full training set)	(92.845)	5.4	37.8	23.5	
	3A. G2P + P2P(Ind) (Full training set)	**1037**	**3.6**	**32.8**	**30.9**	
	3B. G2P + P2P(Ind) (development set)	1358	3.6	32.9	30.4	[32.1 – 33.7]
	4. G2P + P2P(Ind)+P2P(Ded) (development set)	1358 + 68	3.6	32.3	29.9	

first names, 24% for the family names and 14% for the toponyms. The corresponding WER scores for the general-purpose G2P were 34%, 37% and 33%. Comparing these figures with those in Table 2 demonstrates that the remaining errors are mainly in the stress assignment for toponyms and in the phonemization for first names; for family names the situation is somewhere in between.

If the P2P converters are allowed to produce two outputs per name, and if one sets the probability threshold to 0.2 times the probability of the best transcription (according to the P2P that is), one gets about 1.6 transcriptions per name on average. If one then counts as errors the percentage of names for which none of the generated transcriptions is correct, the WERs drop to 33% for first names, 26% for family names and 23% for toponyms. Especially the WIR for first names is multiplied by a factor 3. In about 50% of the first name cases where the general-purpose G2P makes an error, one of the two transcriptions produced by the G2P-P2P tandem is already better than the general-purpose transcription. Making these two transcriptions available may seriously speed up the semi-automatic construction of a first name lexicon.

For comparison, the WER achieved by the G2P on standard words (non-names) is typically below 25%. This shows that names are indeed much more difficult to convert than words; it also shows that the cascaded P2P is not able to bring down the WER that is typical for normal words.

A further examination of the remaining errors for the three name categories (in the development sets) shows that almost no systematic errors were overlooked by the P2P's. There were just a couple of errors with a frequency higher than 10.

First names:

- Names ending at the (Frisian) diminutive affix "–ske" missed the /s/ in the transcriptions. This /s/ was present in the output of the standard G2P, so its removal is due to some kind of overgeneralizations by the P2P.

Family names:

- Names with the phoneme sequence /ɛ[1]-nɑ/ (as in Reinartz) have lost their stress mark, which is an error.

Toponyms:

- A number of names starting with /hɛ/ and /ɔ-rɑ/ ("Oranje") have lost their stress mark on this syllable.
- The P2P removes the stress mark on "Oude" in cases where the reference transcriptions does have it. The same holds for stress marks on the word "pastoor", and "bastion".

At a methodological level, the results of our experiments give interesting guidelines for developing a good P2P:

1. Select a development set of about 2500 names.
2. Use the G2P to obtain an initial transcription of 1000 of these names and correct them manually
3. Train a P2P on this corrected set, and transcribe the remaining names in the development set with the resulting G2P-P2P. Correct the output manually.
4. Retrain the P2P and make transcriptions for the development set using the resulting G2P-P2P tandem.
5. Analyze remaining errors from a deductive perspective
6. Identify features related to higher order phenomena and incorporate them in the learning process.
7. Train the new P2P for optimal results

Higher order phenomena such as syllable and affix identities are automatically detected in a extended version of the learning software, meaning that the deductive analysis in steps 5-7 can probably be skipped for many name types, thus saving a lot of manual efforts.

In terms of time effort, the deductive method is, not very surprisingly, much more costly than the inductive method. The compilation of a deductive rule set for toponyms took at least one day for the full training set of 100k entries and at least half a day for the development set of 20k entries. In contrast, the inductive P2P rule set was generated in one hour (full training set).

Future work will involve the goodness of fit of the computed canonical pronunciations to real pronunciations as they are encountered in everyday speech. These type of comparisons will show the actual validity of our work for Automatic Speech Recognition.

6. Acknowledgements

The presented work was carried out in the Autonomata project, granted under the Dutch-Flemish STEVIN[2] program. The project partners are the universities of Ghent, Nijmegen and Utrecht and the companies Nuance and TeleAtlas.

7. References

[1] Bisani M., Ney H. (2003) "Multigram-Based Grapheme-to-Phoneme Conversion for LVCSR", Procs. Interspeech, 933-936.

[2] Black, A., Lenzo, K., Pagel, V. (1998). "Issues in building general letter to sound rules. Procs. ESCA/ COCOSDA workshop on Speech Synthesis (Jenolan Caves), pp. 77-81.

[3] Boula de Mareüil, P.; d'Alessandro, C.; Bailly, G.; Béchet, F.; Garcia, M.; Morel, M.; Prudon, R. and Véronis, J (2005). "Evaluating the pronunciation of proper names by four French grapheme-to-phoneme converters", Procs. Interspeech, Lisbon, 1521-1524.

[4] Bouma, G. (2000). "A finite state and data-oriented method for grapheme to phoneme conversion", Procs. ACL, 303-310

[5] W. Daelemans, J. Zavrel, K. van der Sloot, A.van den Bosch (2004). TiMBL: Tilburg Memory Based Learner, 5.1, Reference Guide. ILK Technical Report 04-02, http://ilk.uvt.nl/downloads/pub/papers/ilk0402.pdf

[6] Damper, R.I, Marchand, Y., Adamson M.J. and Gustafson, K. (1998), "A comparison of letter-to-sound conversion techniques for English Text-to-speech synthesis." In: Proceedings Institute of Acoustics, 20(6).

[7] Font Llitjós, A. and Black, A.W., "Evaluation and collection of proper name pronunciations online", Procs. LREC2002, Gran Canaria, 2002, 247-254.

[8] Kerkhoff, J., Rietveld, T. (1994). "Prosody in Niros with Fonpars and Alfeios." In: de Haan and Oostdijk (Eds.), Procs.Dept. of Language & Speech, Univ. of Nijmegen, Vol.18, 107-119.

[9] Polyakova, T., and Bonafonte, A. (2006) "Using error-driven approach to improve automatic g2p conversion accuracy". In: TC-STAR workshop on SLT, Barcelona.

[10] Stevens, G., and Boothooft, G. (2006), "Autonomata namencorpora en p2p-evaluatie. http://speech.elis.ugent.be/autonomata

[11] Taylor, P. (2005), "Hidden Markov Models for grapheme to phoneme conversion." In: Procs Interspeech 2005, Lisbon, 1973-1976.

[12] Q. Yang, J.-P. Martens, N. Konings, H. van den Heuvel (2006). "Development of a phoneme-to-phoneme (p2p) converter to improve the grapheme-to-phoneme (g2p) conversion of names",. Procs. LREC, Genua, 287-292.

[1] http://speech.elis.ugent.be/autonomata

[2] http://taalunieversum.org/taal/technologie/stevin/