The AUTONOMATA Spoken Names Corpus

Henk van den Heuvel[1], Jean-Pierre Martens[2], Bart D'hoore[3,]

Kristof D'hanens[2], Nanneke Konings[1]

[1]CLST, Radboud University Nijmegen; [2]ELIS, Gent University, Belgium; [3]Nuance, Gent, Belgium

E-mail: H.vandenHeuvel@let.ru.nl

Abstract

In the Autonomata project we have collected a corpus of spoken name utterances with manually corrected phonemic transcriptions of these utterances. The corpus was designed with the intention to become a major resource for the development of automatic speech recognition engines that can achieve a high accuracy on the recognition of person and geographical names spoken in Dutch. The recorded names were selected so as to reveal the major pronunciation variations that a speech recognizer of e.g. a navigation system with speech input is going to be confronted with. This includes native speakers speaking foreign names and *vice versa*.

1. Introduction

Autonomata is a project funded in the Dutch STEVIN program[1]. One of the goals of the project was to collect a large number of spoken name utterances and to provide manually corrected phonemic transcriptions of these utterances. The Autonomata spoken name corpus was designed with the intention to become a major resource for the development of automatic speech recognition engines that can achieve a high accuracy on the recognition of person names and geographical names spoken in Dutch. The names are further subdivided in first names, family names, street names and city names, and they were selected so as to reveal the major pronunciation variations a speech recognizer of e.g. a navigation system with speech input is going to be confronted with. It is known that very important factors in this respect are: (i) the mother tongue of the speaker (native versus non-native speakers of Dutch) and (ii) the origin of the name (is the name of a foreign or a native origin). Thus, the compiled corpus contains recordings of both native and non-native speakers of Dutch and the uttered names are of Dutch as well as of foreign origins. Table 1 shows the number of speakers times the number of names according to these two dimensions.

	Native speakers of Dutch	Non-native speakers of Dutch
Names of Dutch origin	120 x 175	120 x 75
Names of foreign origin	120 x 75	120 x 175

Table 1: Basic setup of the Autonomata name corpus. Each cell contains the number of name utterances (as #speakers times #utterances per speaker)

The corpus consists of two parts: one part is recorded in the Netherlands (NL), another is recorded in Flanders (FL). This way, it should be able to reveal important cross-regional as well as cross-lingual pronunciation phenomena. Such phenomena are highly interesting in terms of academic and industrial research into the relevant factors for pronunciation variation to be accounted for in commercial applications with a high market potential, such as voice operated navigation systems in cars. Such systems will often be operated by non-native users (e.g. foreign visitors renting a car) and a lot of the names in their vocabulary will (partly) be of a foreign origin (e.g. streets named after foreign celebrities)

Since the corpus is intended to support future pronunciation variation research, it is delivered with no less than four broad phonemic transcriptions per name utterance. These transcriptions are: (1) two transcriptions generated by state-of-the-art grapheme-to-phoneme (g2p) converters, (2) one canonical transcription representing a typical (according to a human expert) pronunciation of that name in the region of recording (NL/FL), and last but not least, (3) an auditory verified transcription representing what was actually pronounced by the speaker.
In the subsequent sections, we discuss various aspects of the construction, validation and distribution of the corpus.

2. Speakers

The corpus includes spoken utterances of 240 speakers living in the Netherlands (NL) or in Flanders (FL). The speakers were selected along the following dimensions:

1. Main region: 50% persons living in the Netherlands and 50% living in Flanders
2. Mother tongue: 50% native speakers of Dutch and 50% non-native speakers
3. Dialect region of *native* speakers: 4 dialect regions per main region
4. Mother tongue of *non-native* speakers: 3 mother tongues per main region
5. Speaker age: one third younger than 18
6. Speaker gender: 50% male, 50% female

The original aim was to select non-native speakers that still speak their (foreign) mother tongue at home and that have a level A1, A2 or B1 (CEF standard[2]) for Dutch.

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

[2] http://en.wikipedia.org/wiki/Common_European_Framework_of_Reference_for_Languages

However, the above strategy appeared to be too restrictive given the limited amount of time there was to finish the speaker recruitment, and given the fact that Flemish schools do not work with the CEF standard. Therefore, the level of proficiency in Dutch was only used as a criterion in the initial phase of the recruitment. It was abandoned as soon as it became clear that the cells of the design would be difficult to fill in this way, especially for non-native speakers with an English mother tongue. Nevertheless, whenever the CEF information was available, even if it was not used as a criterion for selecting the speaker, it was recorded and included in the speaker information file.

The 60 non-native speakers in a region were divided into three equally large groups. But since French is obviously an important language in Flanders and far less important in the Netherlands, the division in subgroups has been made differently in the two regions :

- In **Flanders,** speakers with an **English, French** and **Moroccan** (Arabic) mother tongue were selected.
- In the **Netherlands,** speakers with an **English, Turkish** and **Moroccan** (Arabic) mother tongue were selected.

As foreign speakers mostly live in the big cities and as the dialect region they live in is expected to have only a minor influence on their pronunciation, the dialect region was no selection criterion for these speakers. Native speakers on the other hand were divided in groups on the basis of the dialect region they belong to. A person is said to belong to a certain dialect region if s/he has lived in that region between the ages of 3 and 18 and if s/he has not moved out of that region more than three years before the time of the recording. We adopted the same regions that were also used for the creation of the CGN (Spoken Dutch) corpus[3]. The speaker selection criteria altogether resulted in the following speaker categorization table.

Region	**Origin**	**Dialect region**
120 Dutch (50% males)	60 natives	15 West-Dutch
		15 Transitional region
		15 Northern
		15 Southern
	60 non-natives	20 English
		20 Turkish
		20 Moroccan
120 Flemish (50% males)	60 natives	15 Antwerp & Brabant
		15 East-Flemish
		15 West-Femish
		15 Limburg
	60 non-natives	20 English
		20 French
		20 Moroccan

Table 2: Speaker distribution in the spoken name corpus

[3] http://lands.let.ru.nl/cgn/doc_Dutch/topics/version_1.0/metadata/speakers.htm

Different recruitment strategies for native and non-native speakers and for young and adult speakers have been applied.

A. Recruitment of native speakers

Friends and family

The easiest way to find speakers was to contact friends and family members (by phone, mail,…). After the recording was completed, these persons were then asked if they knew other people who would also like to participate.

Through schools and organizations

For the recruitment of young people (12-18 years old) we collaborated with schools and boarding schools. Grown-ups were contacted through social organizations (e.g. Davidsfonds in Flanders).

B. Recruitment of non-native speakers

Language Schools

Several language schools offer Dutch courses for foreigners. They helped us to find non-native speakers with the right level of linguistic skills. In the Netherlands we contacted many ROCs (Regionale Opleidingscentra) and ISKs (Internationale Schakelklassen) where we actually found many speakers. In Flanders we contacted the Language Center (Talencentrum) in Gent.

Organizations

Specific organizations for foreigners (French, English, Turkish as well as Moroccan) were of great help to contact speakers. For English speakers, we also got a lot of cooperation from English pubs. For the recruitment of Moroccan speakers we approached the Mosques. Even with this help it was not so easy to recruit female Moroccan participants.

3. Names and command words

Each speaker was asked to read 250 proper names and 50 command & control words from a computer screen. The command words are the same for every speaker, but in each region, the names read by a speaker are retrieved from a long list of 2500 names. These lists were created independently in each region, meaning that there is only a small overlap between the names in the two long lists. Once created, the long list was subdivided in 10 mutually exclusive short lists, each containing 250 names: 70% names that are typical for the region (NL/FL) and 30% names that are typical for the mother tongues covered by the foreign speakers (10% for each mother tongue). The typical names for a region were further subdivided in 50% frequent and 50% less frequent words.

For the native speakers we used all 10 short lists, meaning that each name is pronounced by 6 native speakers of a region. For the non-native speakers we worked with only 6 short lists in order to achieve that the same name occurs 3 or 4 times in each non-native subgroup (right column of Table 2).

For the Moroccan names, we chose to select only first names and family names because Dutch speakers will only rarely be confronted with Moroccan geographical names. Furthermore, we adopted the French way of writing for Moroccan names. For all other languages we selected 25% first names, 25% family names, 35% street names and 15% town or city names. So, in total there are 50% person names and 50% geographical names. We selected more street names than city names because there are – logically – more streets than cities in a country.

Exonyms were not included; meaning that we selected "Lille" instead of "Rijsel". Acronyms for highways (e.g. E40, A12) were not selected either.

We also took care that all different standard elements like street, drive, avenue… are present in a proportional way.

Since first names and family names naturally go together, it was decided to reorganize the short lists in such a way that a first name and a family name of the same language of origin and the same frequency class (for the typical Dutch names) are combined into one person name. This means that 69 first names + 69 family names gave rise to only 69 person name utterances, and consequently that there are only 250 – 69 = 181 speech files per speaker.

Since it may be interesting to investigate whether speaker-specific pronunciation phenomena can be derived to some extend from a restricted set of adaptation data, it was decided to let every speaker also pronounce a list of 50 words that are often encountered in the context of an application and that reveal a sufficient degree of acoustic variability to make the word utterances also suitable for acoustic model adaptation. A list of 50 such words was delivered by Nuance (see Table 3). It consists of 15 digit sequences and 35 common command and control words.

0 7 9 1	9 0 2 3	sluiten	opnemen	netwerk
3 9 9 4	9 5 6 0	bevestigen	programmeren	infrarood
0 2 8 9	0 1 2 3	controleren	microfoon	instellingen
5 6 9 4	1 6 8 3	help	stop	herhaal
2 3 1 4	7 8 2 6	ga naar	opslaan	opnieuw
7 8 9 0	activeren	aanschakelen	macro	menu
2 2 2 3	annuleren	Nederlands	controlemenu	opties
5 6 7 8	aanpassen	herstarten	status	lijst
9 0 7 4	ga verder	spelling	batterij	Vlaams
3 2 1 5	openen	cijfer	signaalsterkte	Frans

Table 3: command & control words included in the corpus

4. Recording procedure

The speakers were asked to pronounce an item that was displayed in a large font on a computer screen in front of them. Every participant had to read 181 name items (see section 3) and 50 command word items. To simulate the fact that in a real application environment, the user usually has some idea of the name type s/he is going to enter, the participants in our recordings were also given background information about the origin of the names. To that end, the name items were grouped into subcategories: Dutch person names, English person names, Dutch geographical names, etc. and the name category was displayed before the first name of that category was prompted.

For the presentation and recording we used software that is commonly used by Nuance for the collection of comparable speech databases.

The microphone was a Shure Beta 54 WBH54 headset supercardoid electret condenser microphone. A compact 4 Desktop audio mixer from Soundcraft was used as a pre-amplifier. The 80Hz high-pass filter of the audio mixer was inserted in the input path as a means for reducing low frequency background noise that might be present in the room.

The speech was digitized using an external sound card (VXPocket 440) that was plugged into a laptop. The digital recordings were immediately saved on hard disk. The samples were stored in 16 bit linear PCM form in a Microsoft Wave Format. The sample frequency was 22.05 kHz for all recordings. Before and after every signal there is supposed to be at least 0.5 seconds of silence (this instruction was not always followed rigorously).

In Flanders, a large part of the recordings were made **in studios** (especially those of non-native speakers and adult speakers), the rest was made **in schools** (those of young speakers and non-natives who take courses in a language center). Recordings in schools may be corrupted by some background noise and reverberation. In the Netherlands all recordings were made on location, mostly in schools.

5. Annotations

Each name token has an orthographical and four broad phonemic transcriptions (see Introduction). Two transcriptions were automatically generated by the Dutch and Flemish versions of the Nuance g2p, respectively. A hand crafted example transcription that is supposed to represent a typical pronunciation of the name in the region of recording is created by a human expert. Finally, an auditory verified transcription was produced by a person with experience in making phonemic transcriptions of speech recordings. All phonemic transcriptions consist of phonemes (elements of the CGN phoneme set[4]), word boundaries (represented by a space), syllable boundaries (represented by a hyphen) and primary stress markers (represented by a quote in syllable initial position). This means that the automatically generated transcriptions were converted from the Nuance internal format to the CGN format.

Obviously, the first three transcriptions are the same for all utterances of the same name in one region, and as a consequence, they are provided in the name lists, together with the orthography and the type and language of origin of the name.

The auditory verified transcriptions are specific for each utterance, meaning that there is a transcription file per

[4] http://lands.let.ru.nl/cgn/doc_English/topics/version 1.0/formats/text/fon.htm

speech file. This transcription file was made in Praat[5]. The annotator could listen to the utterance as many times as s/he wished, and s/he was asked to modify (if necessary) the example transcription that was displayed above the signal. The modification was done according to rules outlined in a phonemic transcription protocol that is distributed together with the corpus

The example transcriptions of the 2 x 2500 names were made once and for all before the start of the creation of auditory verified transcriptions:

1. They were extracted from the AUTONOMATA lexica that were available to the project partners. These lexica contained plausible (verified by a linguist) Dutch and Flemish transcriptions: dependent on the region (NL/FL), the first Dutch (or Flemish) transcription was selected as the standard transcription.
2. When they could not be found in the AUTONOMATA lexica, transcriptions were made with the best g2p-p2p tandem (Van den Heuvel et al., 2007) that was available at the time, and corrected by a human expert.

For the sake of consistency we choose to work with example transcriptions for all names, even though for foreign names spoken by native Dutch/Flemish speakers and Dutch/Flemish names spoken by foreigners these standard transcriptions do not offer a time gain compared to transcribing from scratch.

6. Corpus validation

The documentation, the speech recordings, the example transcription and the auditory verified transcriptions were subjected to an external evaluation by BAS Services[6] (Munich). All elements were found to comply with the normal quality standards for such type of annotations, and some discovered shortcomings in the original documentation were remedied.

To demonstrate the validity of the example transcriptions, recognition experiments have been set up for the Flemish speaker set with the Belgian Dutch version of the Nuance VoCon 3200 engine (version 2.5), a state of the art recogniser for embedded applications. The test grammar contained 1810 names in a flat list. The error rate dropped from 12.8% to 8.7% (a relative drop of 32%) overall when substituting the phonemic transcriptions generated by the general purpose rule based g2p converter of the VoCon Embedded Development System by the hand crafted example transcriptions delivered with the data set. Zooming in on native speakers only, the error rate is reduced by a relative 60% (from 6.5% to 2.6%). These results suggest that the example transcriptions are of a high quality.

7. Corpus distribution

The corpus is 9GB large and is distributed by the Dutch HLT-agency (TST-centrale)[7]. The corpus has a rich body of documentation. There is a general documentation file describing all aspects of the corpus construction as well as the format and content of all files that constitute the corpus. Among these files are the phonemic transcription protocol (in Dutch) that was used for the creation of the example transcriptions and the auditory verified transcriptions, a translation of that protocol in English and a document (in Dutch) describing the internal validation experiments that were carried out in the course of the corpus construction process. Examples of speech and annotation files can be viewed at the Autonomata website: http://elis.ugent.be/autonomata (click on 'results')

The external validation documents and the response we gave to these documents, e.g. in terms of adaptations we made to the documentation, are also distributed with the corpus.

7. Future work

In a follow-up project (called Autonomata Too) the same consortium that was responsible for creating the corpus will use the corpus as a resource for studying the relations between computed canonical pronunciations and actually observed pronunciations of street names, place names and POI's (Points of Interest) that constitute the target vocabulary of a modern car navigation application. The working hypothesis is that the auditory verified transcriptions in the corpus can reveal generic pronunciation phenomena that can be exploited to enrich the lexicon of the ASR with pronunciation variants that will help to improve the accuracy of automatic name recognition. In particular Autonomata Too will focus on pronunciation variations representing cross-lingual and cross-regional phenomena due to native and non-native speakers of two regions pronouncing names of various language origins. In that context it will also investigate whether the name recognition accuracy can be further raised by including extra phoneme symbols from foreign languages (so-called xeno-phones) in the phonemic transcriptions of the name tokens (in the corpus) and in the pronunciation variants in the lexicon (see e.g. Eklund & Lindstrom, 2001).

8. Acknowledgements

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

9. References

http://elis.ugent.be/autonomata

Eklund, R. Lindström, A. (2001): Xenophones: an investigation of phone set expansion in Swedish and implications for speech recognition and speech synthesis. In *Speech Communication*, 35, pp.81-102.

Van den Heuvel, H.,.Martens, J.P., Konings, N. (2007). G2P conversion of names. What can we do (better)?, In: *Proceedings Interspeech (Antwerp)*, pp.1773-1776.

[5] http://www.praat.org
[6] http://www.phonetik.uni-muenchen.de/Bas/BasValideng.html
[7] http://www.tst.inl.nl/