A first LVCSR system for Luxembourgish, an under-resourced European language

Martine Adda-Decker[1,2] *Lori Lamel*[2] *Gilles Adda*[2]

[1] Laboratoire de Phonétique et Phonologie, UMR 7018, CNRS-Paris 3/Sorbonne Nouvelle, France

[2]Spoken Language Processing Group LIMSI-CNRS, Orsay, France

martine.adda-decker@univ-paris3.fr, {lamel,gilles.adda}@limsi.fr

Abstract

Luxembourgish is embedded in a multilingual context on the divide between Romance and Germanic cultures and remains one of Europe's under-described languages. We describe our efforts in building an large vocabulary ASR system for such a "minority" language (target language: Luxembourgish) without any transcribed audio training data. Instead, acoustic models are derived from major languages (source languages, here German, French and English). Some scientific and technological issues addressed include: (i) how to build acoustic models if no labelled acoustic training data are available for the under-resourced target language? (ii) how to make use of the new system to accelerate resource production for the target language? First ASR results illustrate the accuracy of the various sets of monolingual and multilingual acoustic models and what these suggest concerning language typology issues.

Keywords:Forced alignment; acoustic modeling; multilingual models; Luxembourgish; Germanic languages, Romance languages.

1. Introduction

Luxembourg, a small country of less than 500,000 inhabitants in the center of Western Europe, is composed of about 65% of native inhabitants and 35% of immigrants. The national language, Luxembourgish ("Lëtzebuergesch"), has only been considered as an official language since 1984 and is spoken by natives (Schanen, 2004). The immigrant population generally speaks one of Luxembourg's other official languages: French or German. Recently, English has joined the set of languages of communication, mainly in professional environments.

As pointed out by (Adda-Decker, 2008) and (Krummes, 2006), Luxembourgish should be considered as a partially under-resourced language, due to the fact that the written production remains relatively low, and linguistic knowledge and resources, such as lexica and pronunciation dictionaries, are sparse. Written Luxembourgish is not systematically taught to children in primary school: German is usually the first written language learned, followed by French.

This paper presents the development of a first Luxembourgish large vocabulary continuous speech recognition (LVCSR) system. To the best of our knowledge, there has never been an LVCSR system for this European language. Efforts were put on gathering all required resources and developping missing blocks: written data for word list and language model development, orthographically transcribed speech data to serve as a reference for system evaluation, a phonemic inventory and a pronounciation dictionary for acoustic phone and word models. The proposed system makes use of acoustic models stemming from different major European languages, without making use of Luxembourgish language specific acoustic training data. This first system will serve as a baseline for further improvements, and will allow to address some linguistically oriented questions. (i) how to build acoustic models if no labelled acoustic training data are available for the under-resourced target language? If multiple monolingual acoustic models from different languages are available for transcribing Luxembourgish audio data, is there a clear preference for one of these languages? (ii) how to make use of the new system to accelerate resource production for the target language? These issues may have important implications for acoustic model development for other under-resourced languages.

The next section introduces the phonemic inventory of Luxembourgish and its correspondance with the three source languages (German, French, and English) used as acoustic model seeds. Written and spoken corpora are introduced in section 3. Section 4 presents the development of acoustic models as well as of Luxemburgish language models. Results are given in section 5 for both monolingual and multilingual pooled acoustic models. Finally, section 6 provides a summary of the results and discusses some future challenges for speech technology and linguistic studies of Luxembourgish.

2. Phonemic inventory of Luxembourgish

The adopted Luxembourgish phonemic inventory includes a total of 60 phonemic symbols including 3 extra-phonemic symbols (for silence, breath and hesitations). Table 1 presents a selection of the phonemic inventory together with illustrating examples (see (**?**) for more information on the phonemic inventory of Luxembourgish). Luxembourgish is characterized by a particularly high number of diphthongs. To minimize the phonemic inventory size, we could have chosen to code diphthongs using two consecutive symbols, one for the nucleus and one for the offglide (e.g. the sequence /a/ and /j/ for diphthong a͡i). We prefered, however, the option of coding diphthongs and affricates using specific unique symbols. Given the importance of French imports, nasal vowels were included in the inventory, although they are not required for typical Luxembourgish words. Furthermore, native Luxembourgish makes use of a rather complex set of voiced/unvoiced fricatives.

Table 1: Sample cross-lingual phoneme associations: Lux. target phonemes associated to same or similar (in grey) phonemes in 3 source languages (Fre, Ger, Eng).

Carrier word (Eng)	Lux	Fre	Ger	Eng
ORAL VOWELS				
l**ii**cht (light)	i	i	i	i
sch**ü**tzen (shelter)	ʏ	y	ʏ	ɪ
f**ä**eg (able)	ɛː	ɛ	ɛː	ɛ
DIPHTHONGS				
l**éi**en (to tell lies)	e͡ɪ	e	e	e
l**ou**nen (to hire)	ɔ͡ʊ	o	o	o

Table 2: Phoneme and training information for native and pseudo-Lux. acoustic models for English, French, German model sets and multilingual superset.

Language	*#native phon.*	*#train. #(h)*	*#Lux. phon.*
English	48	150	60
French	37	150	60
German	49	40	60
Superset(E,F,G)	-	340	3x60

3. Text and speech corpora

Text sources consist in the CHAMBER (House of Parliament) debates and to some extend in news channels, such as delivered by the Luxembourgish radio and television broadcast company RTL. These texts have been filtered according to the criterion described in (Adda-Decker, 2008) in order eliminate sentences which are not in Luxembourgish, because of a frequent switch to French or German (especially in the debates). Table 3 gives a summary of the different training and developement texts used for the study (note that the dev text used for the developement of the word list and language models is different from the developement set used for the ASR experiments).

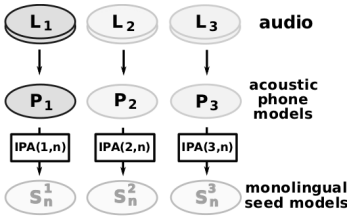
Source	CHAMBER www.chd.lu 2002-2008		RTL www.rtl.lu 2007-2008	
	train	dev	train	dev
Volume	10.10M	104k	0.67M	14k

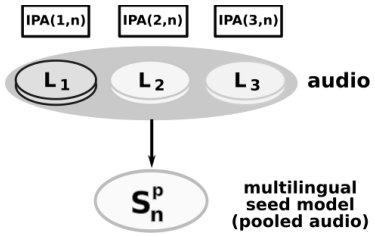
Table 3: "Lëtzebuergesch" training and developement text sources (in number of words).

Beyond large amounts of not yet transcribed audio data, we have 80 minutes of manually transcribed speech from the House of Parliament (*Chamber* debates (70') and from news (10') broadcast by RTL, the Luxembourgish radio and TV broadcast company (Adda-Decker, 2008).

The detailed manual transcripts include all audible speech events, including disfluencies and speech errors.

Figure 1: Acoustic seed models for a target language n (Luxembg.) given phone models P_i of languages L_i ($i = 1, 2, 3$: English, French, German) and IPA symbol correspondances between language i and n IPA(i,n). Top: monolingual S_n^i models. Bottom: pooled multilingual S_n^p model.





4. Acoustic & language models

4.1. Acoustic models

The need to develop acoustic seed models for under-resourced languages has already been addressed in previous research (Schultz, 2001). In the current study, three sets of context-independent acoustic models were built, one for each well-resources seed language (i.e., Engl., Fren., Germ.). The models were trained on manually transcribed audio data (between 40 and 150 hours) from a variety of sources, using language-specific phone sets. The amount of data used to train the native acoustic models and the number of phonemes per language are given in Table 2 (left). Each phone model is a tied-state left-to-right, 3-state CDHMM with Gaussian mixture observation densities (typically containing 64 components). Figure 1 (top) illustrates the development of three sets of pseudo-Luxembourgish acoustic models, each including 60 phones, starting from the English, French and German seed models and mapping the Luxembourgish phonemes to a close equivalent in each of the source model sets (IPA(i,n) in Fig. 1). Table 1 shows a sample of the adopted cross-lingual associations that were used to initialize seed models for Luxembourgish. Some symbols are used several times for different Luxembourgish phonemes. For the diphthongs that are missing in French, phonemes corresponding to the nucleus vowel were chosen. A fourth model set was then formed by concatenating the first three model sets, allowing the decoder to choose among the three models (see Table 2). Finally a set of multiligual acoustic models were trained (see Fig. 1, lower part) using the pooled E,F,G audio data that were labeled using their respective IPA(i,n) correspondances.

4.2. Word list and language models

Given the limited available written data volumes, we decided to limit our word list size to 65k entries, although a larger

vocabulary would certainly be more appropriate for a poorly normalizing and word-compounding language such as Luxembourgish.

A 65 word list was defined from the two training source texts so as to minimize the OOV (out of vocabulary) rate on the CHAMBER dev text according to the method described in (Allauzen, 2004). To this end, two unigram language models were first built using the CHAMBER and RTL training data respectively. Using the CHAMBER development text, the LM perplexity was then minimized by optimizing the interpolation between the two unigram LMs. As a result, the 65k most probable words according to the optimal interpolated unigram LM are kept for the final word list.

The OOV rate of this 65k word list is 2.4% on the CHAMBER dev text, and 6.4% on the RTL dev text. A similar LM interpolation procedure is used to establish the 2, 3 and 4-grams language models. For each order, we have constructed a back-off LM using modified Kneser-Ney smoothing (Chen, 1998; Kneser, 1995), for each source text, and linearly interpolate them so as to minimize the perplexity on the CHAMBER dev text. Figures describing word list and LMs are summarized in Table 4.

Source	CHAMBER dev	RTL dev
OOV (%)	2.4	6.4
2-gram (pp)	162.4	460.3
3-gram (pp)	117.0	406.8
4-gram (pp)	110.7	400.3

Table 4: OOV rates of the 65k word list and 2, 3, 4-grams LM perplexities as measured on the CHAMBER and RTL development texts.

We can see from the figures in Table 4, that the RTL source is poorly modeled by both the word list and the LMs, with an OOV rate and perplexities which are about three times higher than the ones obtained on the CHAMBER source. Due to the difference in size between the two source texts (15 times more words in CHAMBER as compared to RTL), an optimization on the RTL source text gives no improvement on OOV rates and only a slight improvement (355 instead of 406 for a 3-g) on perplexity. Beyond a limited volume for the RTL texts, these data also allow for more writing variants which entail poorer lexical coverage and contribute to higher LM perplexities.

4.3. Pronunciation dictionary

A grapheme-to-phoneme tool has been developed as a PERL script and pronunciation dictionaries have been produced (Adda-Decker, 2008). Table 5 shows a small excerpt of the pronunciation dictionary. Pronunciation development is complex as Luxembourgish spelling rules are permissive and words from different origins follow different pronunciation rules.

4.3.1. Spelling

Lëtzebuergesch spelling standards aim at minimizing pronunciation ambiguities which is in favour of easy pronunciation rules. Concerning Romance or Germanic origins of "Lëtzebuergesch" lexical entries, writing standards may stay more or less close to the language of origin. For Romance items different pronunciation rule sets need to be developed, than for Germanic items. Depending on the origin, `qu` letter sequence of germanic items such as `quälen, quëtschen, Quetschen` calls for a /kw/ pronunciation, whereas Romance rules generally advocate a simple /k/ pronunciation.

LEXICAL ENTRY	PRONUNCIATION
huet (has)	hʊət
lafen (to run)	lafən
héieren (to hear)	heɪərən
dausend (thousand)	dɛwzənt

Table 5: Sample pronunciation dictionary.

4.3.2. Multilingual entries

Lexical entries can be shared by multiple languages as far as they rely on similar alphabets. For short words, combinatorics are reduced and hence many forms can be shared without any etymological link: `ville` means "city" /vil/ in French , and "many" /fɪlə/ in Luxembourgish, `net` means "clear,tidy" /nɛt/ in French, and stands for the negation "not" /nœt/ in Luxembourgish. change" /mɥe/ in French. Among the longer words, shared entries generally imply shared origins. Here one typically finds French or German imports and proper names `Stagiaire, Quartier, Porto, Dubrovnik, Notre-Dame...`.

4.3.3. Variants

Variants concerning the "Lëtzebuergesch" specific phonological process of mobile-n deletion (Krummes, 2006) have been introduced and studied (Snoeren, 2009). French imports may be pronounced according to French standards, or adapted to Luxembourgish, potentially entailing various spellings. Typically the nasal vowel /ã/ changes to /aŋ/, (`Jean`, /ʒã/ becomes `Jang` /ʒaŋ/) and for /õ/ the vowel may become diphthoguized with a nasal coda as /ɔʊn/ in `-tion` words,such as `Abstention, Abstraction, Fonction, Situation...`. A large amount of such imports can be found both in the CHAMBER and in the RTL corpora. Not only the spelling of the vowel can be adapted, but also the French c-letter may be changed to the German k- or z-counterparts `Abstention, Abstentioun; Abstraction, Abstractioun, Abstraktioun`.

Similar to German, Luxembourgish profusely produces compounds. Compounding items from different origins, such as `Beispillfonctioun, Bensinsstatiounen, Wunnengsagglomeratiounen`, are commonly observed in the collected corpora. German imports may be pronounced according to German standards, or adapted to Luxembourgish. Spelling and pronunciation variation here corresponds to items including -ung , which may be written and pronounced either with "u" or with "o" (`Stëmmung, Stëmmong (eng. mood); Meenung, Meenong (eng. opinion)`).

5. Results

First recognition experiments were run on our set of manually transcribed date (70 minutes from CHAMBER and 10 minutes from RTL). Different sets of acoustic models stemming from different languages or from pooled audio data were used. Results are reported in Table 6 in terms of %correct, %substituted, %deleted and %inserted words. The last column gives the word error rate. ASR output words which achieved low acoustic likelihood scores were rejected, resulting in relatively high deletion rates (24.2-30.4%). These rates give an indication of the match/mismatch between the test data and the system's speech models.

We can see that the best results are achieved for the acoustic models stemming from the German audio. German ranks best (54.5% WER, 256 Gaussians) before the pooled models (56.6%). English (62.6%) and French (71.5%) produce significantly higher error rates.

CI models		ASR RESULTS (%)				
source	#G	corr	subs	del	ins	WER
German	64	46.3	29.5	24.2	2.1	55.8
English	64	37.0	32.5	30.4	1.7	64.6
French	64	28.0	34.7	37.3	1.2	73.2
Pooled	64	44.6	27.5	28.0	1.3	56.8
German	256	47.4	28.0	24.7	1.8	**54.5**
English	256	39.1	30.5	30.4	1.7	62.6
French	256	29.8	32.6	37.6	1.2	71.5
Pooled	256	45.0	26.6	28.4	1.6	56.6

Table 6: Recognition results using context-independent (CI) acoustic phone models from German, English, French and Pooled speech data (labeled using the Luxembourgish phonemes). Two sets of models include respectively 64 and 256 Gaussians per state.

Experiments with context-dependent (CD) models are underway.

6. Summary and prospects

The present work focused on the development of a first LVCSR recognition system in Luxembourgish. Luxembourgish language models and a 65k pronunciation dictionary were produced. The issue of producing acoustic seed models for Luxembourgish, a language with strong Germanic and Romance influences was addressed by falling back to related languages' acoustic models. A phonemic inventory was defined for Luxembourgish and linked to inventories from major neighboring languages (German, French and English), using the IPA symbol set. For each of these languages, acoustic seed models were built using either monolingual German, French, English data or multilingual pooled audio.

Our approach to build acoustic models via IPA associations between the Luxembourgish phonemic inventory and those of other languages for which acoustic models are available gives encourageing results. The source language identity of the acoustic models reveals to have a strong influence on the system performance (17% difference between best and worst results). The Luxembourgish speech data are best processed using the German models. English models appear to perform better than the French ones.

The present system, although perfectible along many dimensions, may already be useful for further resource development. A major bottleneck today is the lack of acoustic training data. Manual transcription can be envisioned. However, orthographic standards are only poorly applied by native speakers and manual transcriptions tend to include many writing variants and writing errors. Automatic transcription may be used first to select speech subsets which are relatively easy to transcribe. Second, a speech recognizer produces a normalized transcript, even though it may be more or less correct. An ASR system for Luxembourgish will contribute to produce new resources for this tiny European language, to enable numerous corpus-based studies of spoken and written Luxembourgish and promote Luxembourgish as an e-language. Finally the system may already serve to study pronunciations and acoustic properties of the Luxembourgish sound set.

7. Acknowledgements

This work has been partially financed by OSEO under the QUAERO program.

8. References

Schanen F. (2004). *Parlons Luxembourgeois*, L'Harmattan, 2004.

Adda-Decker M., Pellegrini T., Bilinski E., and Adda G. (2008). "Developments of L'etzebuergesch resources for automatic speech processing and linguistic studies.," in *LREC*, 2008.

Krummes C. (2006). "Sinn si or si si? mobile-n deletion in Luxembourgish," in *Papers in Linguistics from the University of Manchester: Proceedings of the 15th Postgraduate Conference in Linguistics*, Manchester, 2006.

Schultz T. and Waibel A. (2001). "Experiments on cross-language acoustic modeling," in *Proceedings of Eurospeech*, Aalborg, 2001.

Allauzen, A., and Gauvain, J.-L. (2004). "Construction automatique du vocabulaire d'un système de transcription", Journées d'Etude sur la Parole 2004, Fès, 2004.

Chen, S.F., and Goodman J. (1998). "An empirical study of smoothing techniques for language modeling", Technical Report TR-10-98, Center for Research in Computing Technology (Harvard University), August 1998.

Kneser, R. and H. Ney. (1995). "Improved backing-off for m-gram language modeling", Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 1. 1995. pp. 181-184.

Snoeren, N.D., Adda-Decker M. (2009). "Pronunciation and Writing Variants in Luxembourgish : The Case of Mobile N-Deletion in Large Corpora", Proc of 4th Language&Technology Conference, November 6-8, Poznan, Poland, pp. 119-123.