Developments of *Lëtzebuergesch* resources for automatic speech processing and linguistic studies

M. Adda-Decker, T. Pellegrini, E. Bilinski, G. Adda

LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE
{madda,thomas.pellegrini,bilinski,gadda}@limsi.fr

Abstract

In the present contribution we start with an overview of the linguistic situation of Luxembourg. We then describe specificities of spoken and written *Lëtzebuergesch*, with respect to automatic speech processing. Multilingual code-switching and code-mixing, poor writing standardization as compared to languages such as English or French, a large diversity of spoken varieties, together with a limited written production of *Lëtzebuergesch* language contribute to pose many interesting challenges to automatic speech processing both for speech technologies and linguistic studies. Multilingual filtering has been investigated to sort out Luxembourgish from German and French. Word list coverage and language model perplexity results, using sibling resources collected from the WEB, are presented. A phonemic inventory has been adopted for pronunciation dictionary development, a grapheme-phoneme tool has been developed and pronunciation research issues related to the multilingual context are highlighted. Results achieved in resource development allow to envision the realisation of an ASR system.

1. Introduction

The linguistic situation in Luxembourg is challenging for automatic speech processing along at least two dimensions: first *Lëtzebuergesch* is strongly embedded in a multilingual context entailing frequent code-switching and code-mixing. *Lëtzebuergesch* hence represents an interesting testbed for multilingual processing (Adda-Decker and Lamel, 2006). Secondly, *Lëtzebuergesch* may be considered as a partially under-resourced language, as the written production remains low. Such languages presently represent a hot topic in the field of automatic speech processing.

The limited production of written material is related to the easy use of French and German as written communication languages. Further, no orthographic standards were clearly established before the end of the 20th century. This then implies a high degree of variation in the observed written forms. An exhaustive Luxembourgish dictionary was produced after World War II, and this large scale effort actively contributed to the elaboration of spelling standards settled in 1975 and revised in 1999 (Newton, 2002; Schanen and Lulling, 2003). Written Luxembourgish sources, although not very widespread, can yet be found over the last decades and even centuries.

Beyond written material, the existence of sibling resources, providing similar content in both written and audio modalities are particularly helpful for automatic speech recognition (ASR). Steps to an autonomous ASR system include acoustic modeling, pronunciation dictionary and language modeling (Lamel et al., 2002) developments. Most languages make use of broadcast news audio data, together with, as written sources, newspaper texts, news wires and related web pages. In Luxembourg news broadcasts are proposed in *Lëtzebuergesch* on a daily basis, however newspapers are mainly bilingual German/French, with only limited code-switching and code-mixing to Luxembourgish, generally for titles. Yet, it is important to note the recent efforts of establishing word lists and multilingual dictionaries in electronic form (Lulling, 2005). Furthermore concerning the web, *Lëtzebuergesch* actually holds rank 51 in the list of the official Wikipedias under the auspices of the Wikimedia Foundation for various languages (http://meta.wikimedia.org/wiki/List_of_Wikipedias).

For the present study, our first aim was to gather information about existing resources in written and spoken Luxembourgish, which could be helpful to automatic speech alignment and speech transcription system developments. These resources are then examined with respect to ASR needs. In the next section we give some more insight into the linguistic situation in Luxemburg, with a focus on the luxembourgophone situation. Section 3. addresses data collection, and section 4. develops issues in written material processing. The phonemic inventory is presented in 5. and 6. introduces our pronunciation dictionary developments. Section 7. summarizes the achieved results and develops some major challenges for Luxembourgish concerning both speech technologies and linguistic studies.

2. Luxembourg and its linguistic situation

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. *Lëtzebuergesch*, i.e. the Luxembourgish dialect or language, the terminology changes with the questioned linguists, is considered national language of Luxembourg only since 1984. *Lëtzebuergesch* is the (Moselle Franconian) language spoken by native Luxembourgers, French and German being easily used for communication among residents (Schanen, 2004). Major languages practiced by immigrants used to be Portuguese and Italian. The immigrated population generally speaks or learns one of Luxembourg's other official languages: French or German. Recently English has joined the set of prestigious languages of communication and tends to become a major communication tool in professional environments.

The country is often considered a successful example of a multilingual society, however the linguistic situation of Luxembourg is complex. Different reasons contribute to

this: the small size of the country entails a dependence on neighboring countries (Germany, France, Belgium) with a very high rate of cross-boundary exchanges; its historical background and its geographical situation at the frontier of the germanic and romance worlds; and last but not least an important proportion of immigrants of different linguistic origins.

Luxembourg was founded and delimited to its actual size in the first half of the nineteenth century (congress of Vienna 1815, treaty of London 1839) under the pressure of the dominant surrounding nations, rather than as a result of internal independence claims. In parallel, the first half of the 19th century witnessed an important production of *Lëtzebuergesch* literature, with major artists such as Lentz, Rodange and Dicks. At least since that time the question of appropriate spelling conventions arose. German and French used to be the official languages for written administration and communication in Luxembourg since 1848, *Lëtzebuergesch* mainly serving for oral communication. The still young history of Luxembourg has nonetheless actively contributed to the evolution of the Luxembourgish spoken language from the status of a germanic dialect to an autonomous language. Across the evolving political situations of two centuries, several attempts were launched in Luxembourg to establish an orthographic writing system. These successive attempts often relied on opposite prevailing criteria: phonetic precision, morphological considerations, proximity/distance with respect to either German or French conventions. The official orthography, which has finally been adopted in 1975, relies on a mix of criteria, advocating both a relative proximity to German and French, as well as educational and typographical simplicity. These conventions have still recently been revised (1999) to overcome some of the remaining problems and inconsistencies (Moulin, 2005; Schanen and Lulling, 2003).

3. Data collection

Sibling resources, providing both audio and related written material are of major interest for ASR development. The most interesting resource we found here, consists 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.

The Parliament debates are broadcast and made available on the official web site (www.chd.lu), together with written reports (the *Chamber* reports), which correspond to rather close manual transcripts of the oral debates. Another interesting sibling resource stems from the Luxembourgish radio and television broadcast company RTL, which produces news written in *Lëtzebuergesch* on its web site (www.rtl.lu), together with the corresponding audio data. However only very limited amounts of written *Lëtzebuergesch* can be found here, whereas RTL has a profuse audio/video production. Table 1 summarizes the different text and audio resources currently being collected.

12M words have been extracted from the *Chamber* reports (years 2002-2008), which mainly comprise professionally transcribed oral debates. However they also include some written subjects in French. The collected audio data correspond to the debates of the two most recent years, totaling a volume of approximately two hundred hours.

	written	**sibling: audio+written**	
Source:	WIKIPEDIA lb.wikipedia.org	CHAMBER www.chd.lu	RTL www.rtl.lu
Volume:	500k	12M	700k
Years	2008	2002-2008	2007-2008

Table 1: Major *Lëtzebuergesch* text and audio sources for ASR. Collected amounts are given in number of words

4. Written material

Written material is known to be of primary importance to language modeling. However the production in Luxembourgish remains rather limited, as German and French guarantee a larger dissemination. Spelling conventions have been settled only recently, which then entails two drawbacks for the production of written resources: a frequent switch to French or German, even if the scope of dissemination is not an argument, and for the few writers who practise, a relatively high spelling variability in the produced material. Luxembourgish can hence be considered as an under-resourced language, at least from the point of view of written production and of available electronic lexica and dictionaries (Pellegrini and Lamel, 2006).

4.1. Multilingual context

Given the multilingual context in Luxembourg, we have spent some preliminary investigations to measure the number of lexicon entries shared between major European languages (French, German, English, Spanish), and made a comparison with Luxembourgish. For the different languages, word lists typically correspond to the most frequent words occurring in newspapers, news transcripts and possibly some parliamentary debates (Adda-Decker and Lamel, 2006).

In order to give an idea of the volume of lexical entries shared among languages, the number of common entries in the top-n words in recognizer word lists were compared pairwise for the French, Spanish, German and English languages. Results are shown in Figure 1. A word sort by frequency typically puts function words at the top position, followed by general language items, then technical items and finally proper names. If, for the top-50k words, 10k words are shared, this represents 20% of the word list. This proportion is almost achieved for the English-French and the English-Spanish pairs. It is expected that with a higher top-n limit the shared word percentage will increase as the proportion of technical items and proper names becomes larger. Of course shared proportions depend on the language pairs and the type of corpus. For the same type of news corpora used here, English and French share more words than German and Spanish (see Figure 1, left). With a full form comparison, the German language shares the lowest number of entries with the other languages. A 50k word list is not large enough here to include many technical words or proper names, as declension, conjugation and more importantly word compounding produce many distinct general language entries. Fig. 1 (right) compares a Luxembourgish word list, extracted from the parliamentary debates to French, German and English. Curves are relatively similar to the left part of the figure. However there

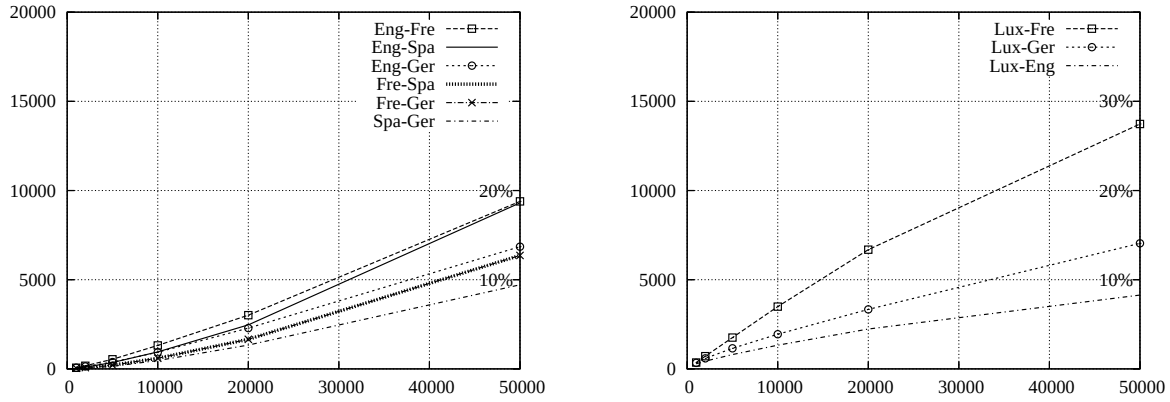

Figure 1: Word list comparisons between pairs of languages. The number of shared words is shown as a function of word list size (including for each language its N most frequent items). **left:** language pairs are among English, French, Spanish and German. **right:** language pairs are Luxembourgish vs French, English and German.

are two noteworthy differences: the percentage of words shared with French is particularly high: French is known to be largely used in administrative and official speech in Luxembourg and there is a frequent code-switching and code-mixing in such contexts. Moreover some parts of the *Chamber* reports are given in French. This may contribute to increase the shared word rates, even though a language filter aimed at removing French texts (see below 4.3.). A second difference (with respect to the left figure) concerns the general slopes of the curves. Curves for Luxembourgish indicate that the contribution of shared frequent words is higher, whereas the part corresponding to proper names remains smaller: the Luxembourgish proper names mostly refer to national personalities, whereas for the other languages broadcast news data include more international proper names. Concerning shared words, some frequent words with common orthography in French and English are for example but, or, son, me, mine, met, as, fond, sale, sort, note, type, charge, moment, service, occasion. Similarly shared words between French and Luxembourgish include an, de, net, et, en, merci, national. Shared entries may be identical only in their surface forms, or may also share (some of) the meaning. Words with some shared meaning are me, charge, moment, type, service, occasion concerning the French-English pair and merci, national for French-Luxembourgish. Yet others completely differ: the word sale in French means dirty, the equivalent of the English sale being soldes, the French word son means his, the English to French translation of son being fils. Similarly most of the examples shown for *Lëtzebuergesch* and French, have completely distinct meanings.

4.2. Text preprocessing

Lëtzebuergesch text preprocessing steps include some language-independent steps, such as raw text extraction, sentence segmentation and de-hyphenation. Depending on the sources, multilingual filtering may become of interest: the CHAMBER corpus includes some French data and potentially also some German. The collected RTL corpus however can be considered as monolingual. Further steps include punctuation normalization and digit conversion.

4.3. Multilingual filtering

An optional language filter may be included in the text preprocessing chain. This filter aims at getting rid of French and German contributions, the scope of which go beyond some local code-switching. The implemented language filter simply relies on language-specific word list, defined as the 65k most frequent words from language-specific training data (typically several hundred million French and German news data, ten million Luxembourgish parliament data). The filtering is carried out on a sentence basis: each word of a sentence is language-tagged, using the three word lists. For a given sentence the selected language tag corresponds to a weighted majority voting. Different weights have been experimented with. In the following, the language tags L, F and G stand for Luxembourgish, French and German respectively.

Raw no language filtering: standard text normalization including hyphenation, sentence and punctuation processing and digit conversions.

F1-1 language filtering using (1,1) weights: a sentence is tagged as Luxembourgish, if the number of L-labeled words is higher than the number of F- and G-labelled words.

F2-1 language filtering using (2,1) weights: a sentence is tagged as Luxembourgish, if the number of L-labeled words is at least twice as high as the number of F-words and simply higher than G-tagged words.

F2-2 language filtering using (2,2) weights: a sentence is tagged as Luxembourgish, if the number of L-labeled words is at least twice as high as the number of F- and twice as high as G-labelled words.

All conditions include the processing steps carried out for the **Raw** format.
The impact of the language filter can be measured using the amount of data rejected. The rejected sentences have been labelled as either French or German. The validity of the approach was checked, by scanning random excerpts of

the rejected data. For higher weights of the filter, the proportion of falsely reject Luxembourgish sentences as either French or German tended to increase. Table 2 gives some figures of the types and tokens included in both the RTL and CHAMBER corpora. For the latter, the multilingual filtering reduces the corpus size by 16 to 22% depending on filter. As expected the proportion of rejections is much higher for French than for German.

filter	corpus	types (k)	tokens (M)	%reject
Raw	RTL	54	0.7	-
Raw	CHAMBER	174	12.1	-
F1-1	CHAMBER	149	10.1	16
F2-1	CHAMBER	141	9.6	20
F2-2	CHAMBER	138	9.4	22

Table 2: Corpus characteristics (types, tokens) for RTL and CHAMBER corpora and for different multilingual filter conditions. The relative rate of rejected data is given with respect to `Raw` counts.

4.4. Word lists and coverage

The written material has been divided into training and test data. Concerning the CHAMBER (12M raw words), most recent debates (2007-2008) have been held out for development and test (approximately 100k words per condition), all earlier contributions (2002-2007) are used for word list and language model development. Concerning the RTL corpus (700k words), small development and test sets (of 10k words each) have been put aside.

Word lists need to achieve high lexical coverage. In the following we wanted to investigate lexical coverage of word lists stemming either from raw (potentially multilingual) or filtered (ideally monolingual) data.

Fig. 2 displays out of vocabulary (OOV) word rates as achieved on CHAMBER training data and development data in different conditions with respect to multilingual filtering. The left figure shows OOV rates measured on raw CHAMBER data (either training or development) using different types of word lists (resulting from either raw or filtered training material). The resulting OOV curves as a function of word list size, inform about the impact of filtering on the word list's lexical coverage capacity. As expected OOV rates are lowest on training data together with a raw word list. The difference in lexical coverage between raw training and development data, increases with the word list size: around 0.6% for a 20k vocabulary, it is more than 1% on the maximum size word list. For the raw dev data, the OOV rate is about 2% for a 60k word list. The remaining curves correspond to word lists established using filtered training data. They show that filtering (i.e. rejecting multilingual items) has a very negative impact on lexical coverage, as long as the input text corresponds to raw data. The right part of Fig. 2 shows similar measures, but here in matched conditions: the filter applied to the text used for OOV measures is the same as the filter used to define the word lists. Curves are shown only for the CHAMBER development data. The curve concerning raw dev data is also displayed in the left part of the figure. However here, all the filtered condition curves feature lower OOV rates than the one from raw conditions, which corresponds to expectations.

Fig. 3 gives OOV measures on the RTL corpus, which is very small as compared to standard ASR system developments. For this condition word lists are built either from RTL only (to get an idea of the impact of too small corpora) or from RTL and CHAMBER interpolated training material. For RTL we change from parliamentary debates to news data. We can see first that the word list merely stemming from RTL data, remains too small, entailing very high OOV rates on the development data. Interpolation with a large corpus, even if different in type, contributes to lower OOV rates, for fixed vocabulary size conditions. These can be even further reduced thanks to larger possible vocabularies. The overall slope of the curves are somewhat chaotic in the central part (around 100K), and more in-depth analysis are required to elaborate some explanations. A global comment is that OOV rates are much higher here than for the CHAMBER, rising above 4% in all dev conditions. Many words, in particular proper names, appearing in news data, are missing in the CHAMBER corpus.

Concerning the composition of the *Lëtzebuergesch* word list, there are very few French or German entries among the 10k most frequent items. However checking the word list of the Chamber debates, we can note for example a high proportion of French import verbs, mainly used for technical and specialized domains. `abordéieren`, `aboutisséieren`, `absolvéieren`, `absorbéieren`, `accordéieren`, `agéieren`, `élaboréieren`, `intervenéieren`, `irritéieren`, `clarifiéieren`, whereas the verbs of German origin are used here, can easily used in more every day situations and vernacular language without appearing to be pedantic. `abezéien`, `aféieren`, `aschätzen`, `aschreiwen`, `ausgläichen`, `héieren`, `kloerstellen`, The same meaning may be given either by a word of either Romance or German origin, (e.g. `clarifiéieren` or `kloerstellen`), the choice may contribute to fix the language style.

4.5. Language models

Given the limited volume of training data, and the OOV rates of the different word lists, we have selected 60k and 100k as potentially interesting vocabulary sizes. However, given the linguistic characteristics of Luxembourgish, 300k vocabularies are certainly recommended, provided large enough corpora are available (McTait and Adda-Decker, 2005).

A series of language models have been estimated for 60k and 100k word lists and different multilingual filtering conditions. Corpus interpolation has optionally been carried out, in order to measure potential improvements in language model perplexity on the different development sets.

Table 3 gives some representative perplexity values for the different conditions. Only a small subset of the results is presented; for instance we have kept only one filtering condition (F1_1), as the variations among the different filterings are not meaningful.

Perplexities have been measured without rejecting OOV items, by relying on a virtual vocabulary size of 2M words. This results in very small probability masses for unknown

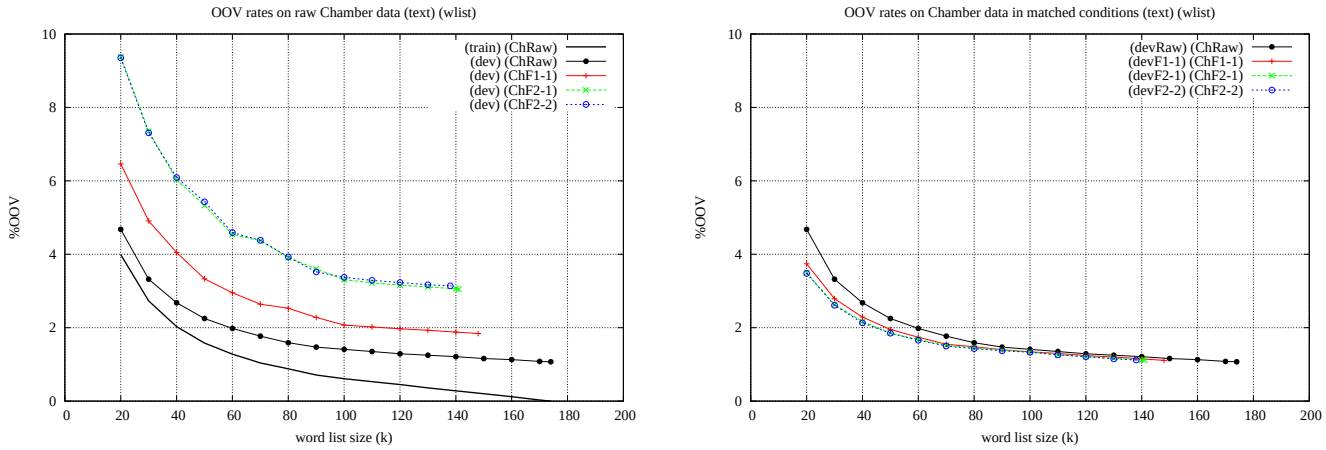

Figure 2: Out of vocabulary (OOV) word rates measured for different word lists on CHAMBER training and development data. Conditions specify the data used for OOV measures (train/dev) followed by the word list type (ChRaw, ChF1_1...).

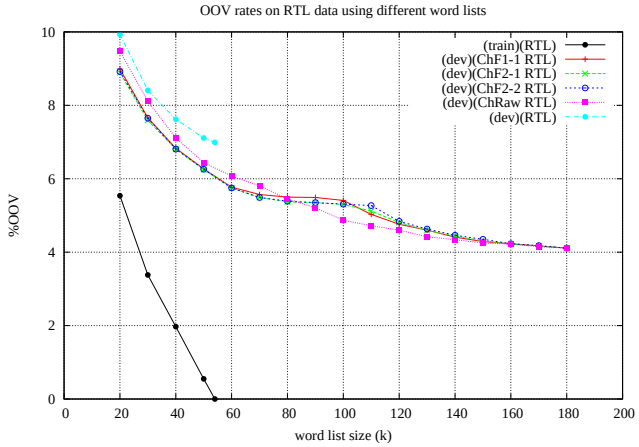

Figure 3: Out of vocabulary (OOV) word rates measured for different word lists on RTL training and development data.

events. This method is the best compromise to handle the OOV problem during the perplexity computation. Other classical calculations result in an artificial decrease of perplexity with increasing OOV rates. Concerning the RTL data, interpolation with the Chamber data is required to increase word list sizes beyond the intrinsic vocabulary of the small RTL corpus (54k). For the 60k vocabulary, interpolation allows only to add few CHAMBER-specific lexical entries. In this condition, we can see that multilingual filtering might be of interest for both the word list and the LM, as the smallest ppx is achieved (498), although the OOV rate (5.8) is higher than in the equivalent 100k condition. However overall variations are small and perplexities are very high in all conditions, reflecting a lack of appropriate training data.

corpus	wlist size	wl filter	txt	OOV	ppx
RTL	100k int	raw	raw	4.9%	513
RTL	100k int	F1_1	filt	5.3%	508
RTL	60k int	raw	raw	6.1%	512
RTL	60k int	F1_1	filt	5.8%	498
CHAM	100k	raw	raw	1.4%	105
CHAM	100k	F1_1	raw	3.3%	121

Table 3: Trigram language model (LM) perplexity (ppx) measures on RTL and CHAMBER development data. Word lists and LMs may be interpolated (int).

Perplexities on the CHAMBER data feature relatively small values close to 100. In this condition the amount of training data available seems satisfactory, even though an order of magnitude of additional data could certainly be of help.

5. Phonemic inventory

The word lists derived from the written material allow to fix optimal vocabularies for the ASR system. A further step consists in providing pronunciations for each lexical entry. Such pronunciations rely on a phonemic inventory. Hereafter we will give details about the the *Lëtzebuergesch* phonemic inventory, detailing vowels, diphthongs and consonants (Schanen, 2004).

The *Lëtzebuergesch* phonemic inventory is characterized by a particularly high number of diphthongs. Concerning linguistic studies (Moulin, 2005), many aspects of

the Luxembourgish language have been explored on limited spoken material. They still need to be investigated on a larger scale and on fluent speech, in particular for pronunciation variants, including phonological phenomena such as the *mobile-n deletion*, also known as Eifeler Regel (Krummes, 2006). The existing phonetic, phonological, prosodic, lexical and morpho-syntactic studies are generally carried out using limited objective observations. Large oral corpus-based studies might be carried out, provided *Lëtzebuergesch* automatic speech alignment and transcription systems were available.

IPA	carrier word
i	liicht, driibseg
ɪ	Lidd, Hiwwel
y	Süden, üben, Äppeljus,
ʏ	schützen, Büro
e	Leed, bereet
ɛ	drécken, Réck, zéng
ɛː	fäeg, Här
ɛ	fetteg, hätt
a	laachen, hat, gebak, aacht
ɐ	Lach, hatt, Papp, Mamm
o	loossen, Rot, Joren
ɔ	Loft, hoffen
u	Luucht, uzen
ʊ	luppen, huppen
ø	Föhn, Hörer, Milieu
œ	lëften, hëllefen, net
ə	fuddelen, elo, esou, et

Table 4: Vowels of *Lëtzebuergesch*.

6. Pronunciation dictionary

In the following we raise some issues concerning high-quality pronunciation dictionaries.

6.1. Spelling

Lëtzebuergesch spelling standards aim at minimizing pronunciation ambiguities, even though minor problems remain. For example the au letter sequence is ambiguous with respect to /ɛʊ/ (Haut) or /ɐʊ/ (haut) pronunciations. Concerning Romance or Germanic origins of *Lëtzebuergesch* lexical entries, writing standards may stay more or less close to the language of origin, as discussed in section 2. For French words such as attaquer (eng. to attack) or abdiquer (eng. to abdicate), the corresponding *lëtzebuergesch* orthographic forms are attackéieren and abdiquéieren, after the official Luxembourgish CORTINA spellchecker (cortina.lippmann.lu). For Romance items different pronunciation rule sets need to be developed, than for Germanic or Moselle-Franconian 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.

6.2. 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, muer /mʊə/ is the *lëtzebuergesch* word for "tomorrow", and stands for "to slough, to 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, Taxe, Projet, Gesellschaft, minimal, Berlusconi, Blair, Kohl, Wolfowitz, Porto, Dubrovnik, Notre-Dame....

IPA	carrier word
p	**p**aken, ra**pp**en, o**p**
b	**b**aken, erla**b**en
t	**t**aasch**t**en, ha**t**en, Ro**t**
d	**d**roen, lau**d**en
k	**k**achen, ba**k**en, Volle**k**
g	**g**oen, u**g**ebak, a**g**educkelt
ʦ	**z**apen, **Z**äit, schwät**z**en, Sa**z**
m	**M**a**mm**, **m**aachen, ë**mm**er, Ha**m**
n	**N**aupen, **n**uets, **N**éidesch, ä**nn**eren, hu**nn**,
ŋ	Aa**ng**el, me**ng**en,
f	**F**eier, **F**irle**f**anz, Ofla**f**, **v**éier, a**v**erstan
v	Wieder
s	**C**entre, Chip**s**en, Kla**ss**, lue**s**
z	**S**ummer, A**s**az
ʃ	Schoul
ʒ	**G**ilet, **J**ang, In**g**enieur
ç	Ee**ch**en, Ki**ch**en, Spii**ch**ten, te**ch**nesch, alldeegle**ch**
j	ee**g**en, Li**g**ener, aartele**g**, alldee**g**lech
h	**h**aut, **h**ei, Äis**h**elleg, un**h**alen
x	aa**ch**t, Zu**ch**, fa**ch**en, Mëttwo**ch**
ʁ	**R**ou, **r**abbeleg, k**r**ibbelen, u**r**uffen, Rack
l	**l**uewen, **L**eit, Bël**l**s, eide**l**, gehoⅼⅼ
w	**W**ebmaster
j	**j**o
ɐ̯	Kann**er**
n̩	moos**sen**
m̩	fein**em**
l̩	Wue**rel**

Table 5: Consonants of *Lëtzebuergesch*.

IPA	carrier word
eɪ	léieren, héich
ɛɪ	läit, fräi, Zäit
ɑɪ	Leit, leien, dreiwen, Haiser
ɔɪ	Europa, Rheuma, moien
ɛʊ	lauschteren, Haut (eng. skin)
ɑʊ	lauden, haut (eng. today)
ɔʊ	lounen, Houscht
ɪə	liesen, hien
ʊə	lueden, huet

Table 6: Diphthongs of *Lëtzebuergesch*.

6.3. Variants

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; Emancipatioun, Emanzipation, Emanzipatioun.

Similar to German, Luxembourgish profusely produces compounds. Compounding items from different origins, such as Beispillfonctioun, Bensinsstatiounen, Investitiounsverloscht, Welt-Health-organisatioun, Wunnengsagglomeratiounen, are commonly observed in the collected corpora. German imports may be pronounced according to German standards, or adapted to Luxembourgish. A major source of 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)). Other items are used with a fixed spelling/pronunciation: e.g. in Wartesall /vʌrtəzʌl/ (eng. waiting-room), the first item Warte- strictly follows the German spelling and pronunciation, whereas other similar compounds admit either both German and *Lëtzebuergesch* variants (e.g. Waardelëschten, Wartelëschten) or just the *Lëtzebuergesch* form (e.g. Waardezeiten).

6.4. Pronunciation rules

A grapheme-to-phoneme tool has been developed as a PERL script and pronunciation dictionaries have been produced. An important issue here remains the proper determination of the origin(s) of a given word (Luxembourgish, German, French, English...), as spelling rules are partially inherited from the language of origin. For example the letter "g" is ambiguous with respect to graphemic context and language of origin: Antigel follows French grapheme-to-phoneme rules : g | {i}_{e} ⇒ /ʒ/, whereas Aangel follows German/Luxembourgish conventions : g | {n}_{e} ⇒ /ŋ/, agelaf g | {a}_{e} ⇒ /g/. The script is presently composed of a set of rules, mainly addressing the Luxembourgish spelling rules and some major exceptions. Language-specific rule sets need further developments. *Lëtzebuergesch*, and more generally languages in multilingual contexts, introduce new challenges to pronunciation dictionary design and development. These partly meet the challenges of multilingual speech processing.

7. Summary and prospects

In the present contribution the complex linguistic situation in Luxembourg has been briefly described. For ASR development sibling resources, providing similar content in both written and audio modalities are particularly helpful. A corpus including news and parliamentary debates has been collected. A text preprocessing and normalization chain including multilingual filtering has been defined corresponding to *Lëtzebuergesch* specificities. An important amount of mainly French, but also German imports can be found in the word lists. Lexical coverage and language model perplexity measures have been carried out.

The available CHAMBER data allow to achieve relatively low perplexities, promising reasonable automatic transcription results for the future. The RTL news corpus, with perplexities around 500, certainly requires additional appropriate training material. However, the already achieved results in resource development allow to envision the realisation of an ASR system. An important side-effect of ASR systems is the production of annotated speech corpora and a large panel of corpus-based studies can be carried out in order to improve our knowledge of *Lëtzebuergesch* and to examine the multilingual reality in Luxembourg for different communication situations.

Acknowledgments

The authors would like to thank the members of the CPLL (Permanent Council of the Luxembourgish Language) for their help with sibling corpora and fruitful discussions. This work has been partially financed by OSEO under the Quaero program.

8. References

M. Adda-Decker and L. Lamel. 2006. *Multilingual pronunciation dictionaries in Multilingual Speech Processing*. Elsevier.

C. Krummes. 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.

L. Lamel, J.L. Gauvain, and G. Adda. 2002. Lightly supervised and unsupervised acoustic model training. *Computer Speech and Language*, 16(1):115–229.

J. Lulling. 2005. *Luxdico - dictionnaire bilingue luxembourgeois - français*. Presses universitaires de Namur.

K. McTait, M. Adda-Decker. 2003. The 300k LIMSI German Broadcast News Transcription System. *In Proceedings of Eurospeech, Geneva September 2003.*

C. Moulin. 2005. *Perspektiven einer linguistischen Luxemburgistik - Studien zu Diachronie und Synchronie*. Universitätsverlag WINTER Heidelberg.

G. Newton. 2002. *Studies from the Germanic Languages - The standardization of Luxembourgish*. John Benjamins Publishing Company.

T. Pellegrini and L. Lamel. 2006. Experimental detection of vowel pronunciation variants in amharic. In *LREC*, Genoa.

F. Schanen and J. Lulling. 2003. Introduction à l'orthographe luxembourgeoise. In *www.cpll.lu/ortholuxs_l.html*, G.-D. de Luxembourg.

F. Schanen. 2004. *Parlons Luxembourgeois*. L'Harmattan.