The study of writing variants in an under-resourced language: Some evidence from Mobile N-Deletion in Luxembourgish

Natalie D. Snoeren, Martine Adda-Decker, Gilles Adda

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

Abstract

The national language of the Grand-Duchy of Luxembourg, Luxembourgish, has often been characterized as one of Europe's under-described and under-resourced languages. Because of a limited written production of Luxembourgish, poorly observed writing standardization (as compared to other languages such as English and French) and a large diversity of spoken varieties, the study of Luxembourgish poses many interesting challenges to automatic speech processing studies as well as to linguistic enquiries. In the present paper, we make use of large corpora to focus on typical writing and derived pronunciation variants in Luxembourgish, elicited by mobile -n deletion (hereafter shortened to MND). Using transcriptions from the House of Parliament debates and 10k words from news reports, we examine the reality of MND variants in written transcripts of speech. The goal of this study is manyfold: quantify the potential of variation due to MND in written Luxembourgish, check the mandatory status of the MND rule and discuss the arising problems for automatic spoken Luxembourgish processing.

1. Introduction

The national language of the Grand-Duchy of Luxembourg, Luxembourgish or "Lëtzebuergesch", has often been characterized as one of Europe's under-described and under-resourced languages. Just like the English language, Luxembourgish can be considered as a mixed language with strong Germanic and Romance influences. It is hard to estimate the precise proportion of Germanic and Romance influenced words in Luxembourgish, as these proportions are modulated by the communicative settings. For instance, although vernacular Luxembourgish is mainly influenced by Germanic stems, there are Romance words to be found as well (*Merci*, *Äddi*,"Adieu", *futti*, colloquial of the French "foutu" meaning "damned"). Nonetheless, more technical and administrative communication systems include a high proportion of Romance words (*Konditioun*, "condition"; *agéieren*, "to act"; *abordéieren*, "get into").
It is estimated that about 300,000 people worldwide speak Luxembourgish. As was previously pointed out (Adda-Decker et al., 2008; Krummes, 2006), Luxembourgish should be considered as a partially under-resourced language, mainly because of the fact that written production remains relatively low. Rather surprisingly, written Luxembourgish is not systematically taught to children in primary school, German being usually the first written language learned, immediately followed by French (Berg and Weis, 2005). Although many efforts have been made in the past to standardize an official orthography of Luxembourgish, no officially recognized spelling system was being recognized until the adoption of the "OLO" (*ofizjel lezebuurjer ortografi*) in 1946, which aimed at producing written forms that clearly diverge from German orthography. In spite of its official character, it never became popular in schools. A more successful standardization eventually emerged from the work of specialists charged with the task of creating a dictionary that was published between 1950 and 1977 (Linden, 1950). Nonetheless, up until today, German and French are the most practiced languages for written administrative purposes and communication in Luxembourg, guaranteeing a larger dissemination, whereas Lëtzebuergesch is the main language used for oral communication purposes between native speakers of Luxembourgish.
The strong influence of both German and French, among other factors, can explain the fact that Luxembourgish exhibits a large amount of both pronunciation and derived potential writing variants. For instance, it is fairly common to have several regional pronunciations for function words (e.g. the English personal determiner "our" can be written and pronounced as *eis* [ajs], *ons* [ɔns], *is* [i:s]). These pronunciation variants may give rise to resulting variations in written Luxembourgish, as Luxembourgish orthography strives for phonetic accuracy (Schanen, 2004). The question then arises, in particular for oral transcripts, whether the written form reflects the perceived pronunciation form or whether some sort of normalization process is at work that eliminates part of the variation. With respect to automatic speech recognition, text normalization is an important issue in order to achieve reliable estimates for n-gram based language models, and even more so for poorly resourced languages. The limited production of written material is related to the fact that French and German are used as the two main written communication languages. Apart from written materials, the use of sibling resources that provide similar content in both written and auditory modalities has proven to be particular helpful for automatic speech recognition (ASR). In Luxembourg, news broadcasts are delivered in Luxembourgish on a daily basis. Newspapers, however, remain for the most part bilingual German/French with occasional code-switching to Luxembourgish (especially for titles). In spite of the ubiquitous influence of German and French on Luxembourgish, a lot of effort has been made over the past few years to establish Luxembourgish word lists and multilingual dictionaries in electronic form (Lulling and Schanen, 2009). As far as web resources are concerned, Luxembourgish holds rank 55 in the list of 272 official wikipedias (cf. the Wikimedia foundation for various languages). This means that about 28000 Wikipedia articles have been created in the Luxembourgish language,

showing that there is a societal demand to communicate in the Luxembourgish language.

2. ASR and the study of variants

2.1. Dealing with variants

Over the past decades, one of the main challenges in automatic speech recognition pertained to the question as to how to handle variation (Strik and Cucchiarini, 1999). Typically, written variants are being dealt with through text normalization processes. Two differently represented variants may refer to the same meaning, so instead of treating these as different, one can treat them as instances of the same underlying sequence. Ultimately, the goal of text normalization is to remove "noise", achieve better lexical coverage and more precise language models that are critical to the development of performing ASR systems. In text normalization, one defines the limits of what will be a *word* in the system. There is, however, an apparent contradiction that needs to be resolved during the optimization of both lexical coverage and language model precision. On the one hand, a minimal number of variants is required so as to reduce the number of Out Of Vocabulary (OOV) words (i.e. the words of the texts that are not part of the vocabulary). On the other hand, one needs to limit the occurrence of ambiguities in order to increase the precision of the language model. Text normalization (mainly language-dependent) is the result of the trade-off between these two conflicting criteria (Adda et al., 1997).

The need for modeling pronunciation variation stems from the simple fact that the words of a language are pronounced in many different ways due to variations in speaking style, interlocutor, communicative context, accent or dialect, socio-economic factors and so forth. Indeed, pronouncing words implies that *they are strung together into connected speech* (Kaisse, 1985) as opposed to the pronunciation of isolated words. As a consequence, all sorts of interactions may take place between words in connected speech, which will result in the application of many phonologically motivated variations such as assimilation, co-articulation, segment reduction, insertion, and deletion. One means to deal with variation that occurs in word pronunciation, is through the creation of specific lexica that incorporate the most commonly observed phonological variants for each word in the lexicon. However, it has previously been shown that simply adding pronunciation variants at the lexical level does not suffice to obtain the best recognition performances (Riley and Ljolje, 1995). Better results are generally obtained when the probabilities of the pronunciation variants are equally taken into consideration, either in the lexicon or in the language model (Strik and Cucchiarini, 1999). The commonly adopted acoustic HMM (Hidden Markov Model) structure can implicitly account for some amount of speech lengthening, especially stemming from hesitation phenomena, and for parallel variants (Adda-Decker and Lamel, 1999). However, pronunciations with a number of phonemes differing from the one specified in the pronunciation dictionary are generally poorly dealt with (Greenberg, 1999). Given the specificities of Luxembourg, it appears important to check the variations arising from the different languages in contact in Luxembourg. One can then focus on Luxembourgish-specific phonological phenomena, such as mobile n-deletion (hereafter shortened to MND, following Krummes (2006), also known as the Eifeler rule (Gilles, 2005; Schanen and Lulling, 2003).

2.2. Effect of MND on written and pronunciation variants in Luxembourgish

According to the phonological rule of MND, a word-final -n is only retained before a vowel or before one of the following phonemes: {n, d, t, ts [z], h}. Any other phonemic right contexts cause the deletion of the final -n. The phoneme -n can also be deleted within compound-word boundaries. That is, the first element of compound words ending in -n generally undergoes MND. So, for instance, given a first element of the word *Fritten* ("French fries"), the -n is preserved before /d/ as in *Frittendëppen* ("chip pan"), but generally deleted before /f/ as in *Frittefett* ("frying fat"). Prefixes ending in -n, also undergo MND. Given the preposition *an* ("in"), prefixed to the verb *droen* (Ger. "tragen", Eng. "to carry") results in *androen* ("to register"), whereas prefixed to a word such as *Fett* (Fr. "gras", Eng. "fat"), results in the verb *afetten* ("to grease").

In the current contribution, we propose to investigate written and pronunciation variants in Luxembourgish that are elicited due to MND, by looking into large transcribed corpora (Adda-Decker, Pellegrini, Bilinski, & Adda, 2008), i.e. manual transcriptions of recorded speech from either the Chamber debates or web news reports. By doing so, we are in an excellent position to characterize this particular variant and to establish with what kinds of variants the Luxembourgish listener is actually confronted with.

3. The current study

3.1. Data selection

Sibling resources that provide both audio and corresponding written materials are of major interest for ASR development. The most interesting resource we have come across until so far for Luxembourgish, consists of the *Chamber* debates (House of Parliament) and to a lesser extent news channels that are delivered by the Luxembourgish radio and television broadcast company. The Parliament debates are broadcast and made available on the official web site (www.chd.lu), together with written *Chamber* reports, that correspond to fairly reliable manual transcripts of the oral debates. Another interesting sibling resource stems from the Luxembourgish radio and television broadcast company RTL, that produces news written in Luxembourgish on its web site (www.rtl.lu), together with the corresponding audio data. However, it must be noted that only a very limited amount of written Luxembourgish can be found here, whereas RTL has a profuse audio/video production. Table 1 summarizes the different text and audio resources that are currently being collected for further analysis.

3.2. Characterizing potential mobile -n sites

As was mentioned before, MND concerns the deletion of a word-final -n, giving rise to a variant of the same lexical item. Following the official Luxembourgish orthography, Luxembourgish words such as *wann* and *wa* ("when") are both recognized as existing lexical items and, as such, listed in the dictionary. Because of the fact that our corpora contain items that can occur without word-final -n, with -n,

	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 Luxembourgish text and audio sources for ASR studies. Collected amounts are given in word numbers, adapted from Adda-Decker et al., 2008.

or double -n, we first sought to know how many Luxembourgish word-final -n (or -nn) words also occur without a word-final -n (or -nn). These items correspond to potential MND sites. To this end, an extraction tool was developed and implemented that took as input the word list derived from the word tokens of the corpora and produced as output a compressed word list merging all the word-final -n variants in the format of the annotation that list word-final -n (or -nn) items that also exist without -n. A few examples are given below:
[1] *gezwonge#n* ⇒ *gezwonge*; *gezwongen* (Eng. "forced");
[2] *ausgi#nn* ⇒ *ausgi*; *ausginn* (Eng. "spent");
[3] *si#n#nn* ⇒ *si*; *sin*; *sinn* (Eng. "are").
The input word list from the transcriptions includes 194k distinct word forms. The correct orthography of these words can be checked using the official Luxembourgish spelling checker developed by the Centre de Recherche Public G. Lippmann with the support of the CPLL (*Conseil Permanent pour la Langue Luxembourgeoise*). This checking allows to list all the words that are considered to be officially admissible Luxembourgish word forms. This officially correct list is termed here the Cortina list and includes 121k words. As such, the word list can be thought of as a standardized type of dictionary, contrary to the word lists that are derived from the transcriptions. Since the input word list concerns high-quality transcriptions, the size difference between the input word list and the Cortina list cannot simply be attributed to transcription errors. Moreover, a lot of Luxembourgish lexical entries have been attested that are not listed in Cortina such as a number of compound words (e.g., *Babyjoren*; *Bäckermeeschter*), acronyms (*NATO*), proper names (*Fischbach*), or toponyms (*Guantanamo*).The results of our word-final -n variant merging are summarized in Table 2. The results of the word-final n merging show that a relatively large number of word-final -n items also occur without the final -n, according to the Cortina list (4.9% of the word types). This proportion more than triples in the Transcriptions list (15.6%), which is not surprising as human transcriptions generally allow for more variation, including potential errors. Another issue might be related to the fact that the Cortina spell checker did not include all the possible variants due to MND. The large amount of additional word-final -n variants may arise from genuine variation in the produced speech due to the MND process. In future studies this point will be investigated, in particular by confronting sibling written and oral modalities. Although the number of -#n#nn type items in the Cortina list is very low (136 items), it is interesting to note that this type of variants is virtually not occurring in the transcriptions. One possible explanation might perhaps be related to avoidance of redundancy when transcribing (i.e. two orthographic representations correspond to the same phonetic variant). These raw measurements provide us with some interesting clues about potential mobile -n sites in Luxembourgish. The fact that a lot of the resulting MND variants are already listed in word-lists might be helpful in explaining under what circumstances MND occurs in Luxembourgish speech.

-n variants	Transcriptions	Cortina
-	194k	121k
#n	30318 (15.6)	5894 (4.9)
#nn	583 (0.3)	101 (0.1)
#n#nn	15 (0.0)	136 (0.1)

Table 2: Word type frequencies (%) of potential mobile -n items and variants as found in the lists derived from the transcribed corpora and in the Cortina list (official orthography). The first line indicates the full word list sizes.

3.3. MND in transcriptions

The goal of a second investigation was to find out whether the MND rule is being respected in two transcriptions from the Chamber debates and one transcription from a news channel (transcribed by professional transcribers, who are native speakers of Luxembourgish). A PERL script was implemented that allowed to count the number of lexical items containing a word-final n in the phonemic contexts in which MND occurs. Table 3 gives a summary of the word frequency and respective type frequencies (%) of violation of the MND rule (taking into account the exceptions to the rule such as word-final *-ioun* where word-final -n is always being retained). These numbers suggest that there are relatively few cases for which the MND rule is being violated. MND violations may include nouns followed by prepositions (*Bühn fir*, "stage for", which in this particular example should not be considered as an MND violation but as an MND exception). MND Violations do seem to affect other syntactic categories as well (e.g., Verb-Preposition: *huele#n fir* "take for", *kucke#n vun* "watch from", Determiner-Noun: *de#n Referendum*, "the referendum", Adj-Noun: *anere#n Länner*, "other countries". Further examples include nouns followed by verbs (*Kirchen gét*) which in this case is a genuine MND violation. Obviously, a more in-depth analysis is clearly called for in order to determine whether the number of potential -n sites varies as a function of syntactic and/other linguistic factors.

Transcription:	Ch1 (12395)	Ch2 (1952)	News (2326)
MND viol.:	0.39	0.46	2.53

Table 3: Word token frequencies (%) and MND violation type frequencies (%) for three transcriptions.The first line indicates the full word list size.

Given these observations, the MND rule is fairly well respected and these results make even more sense in the light of the relatively large number of listed variants resulting from MND that was mentioned before. In order to verify this hypothesis, however, the next step would be to collect more linguistic information about the type of items that undergo MND and to see whether this information correlates with the potential mobile -n words that are listed in the dictionaries and recognized as lexical items in their own right by Luxembourgish listeners. Finally, transcriptions need to be checked against oral productions to clarify whether MND is similarly respected in the oral modality.

3.4. MND and word list coverage

Language model development in ASR requires that the word lists that are being used achieve high lexical coverage. As was previously mentioned, text normalization processes are employed to obtain good lexical coverage. Adda-Decker et al. (2008) looked into lexical coverage of Luxembourgish word lists from raw (i.e. potentially multilingual) and filtered (i.e. approximating monolingual) data by using the *Chamber* training and development data. It was found that, concerning the composition of the different word lists, there were actually very few French and German entries in the filtered *Lëtzebuergesch* word list, whilst the word lists from the *Chamber* debates contained a high proportion of Romance import verbs. Following Adda-Decker et al. (2008), we sought to quantify the impact of mobile -n variants on lexical coverage in Luxembourgish. To this end, we used the Chamber corpus that consists of 12M raw words as training data to build different size word lists (i.e. system vocabularies). A held out development set of 100k raw words was then used to measure the percentage of words covered by the different size word lists on the new data. The complementary measure of unknown words, termed Out of Vocabulary (OOV) words, is displayed in Figure 1 as a function of word list size (varying between 10k and 150k lexical items). The corresponding curves inform about the impact of MND, that is, after filtering out all word-final -n items, on the word list's global lexical coverage capacity. As can be seen from the Figure 1, OOV rates overall decrease as the word list size increases. More importantly, the difference between the MND filtering and the standard development data is relatively important at a low word list size. However, the difference between the two curves reduces as the word list size increases (beyond 80k). In light of the observed differences between the MND filtering and development data slopes, it is relevant to see how the curves for word-final phonemes other than -n fare with respect to the development data. Figure 2 displays the curves for the 8 most frequent word-final phonemes other than word-final -n, whereas Figure 3 zooms in on the 20k-80k word list size range. It can be seen from these two Figures that the curves closely parallel the development data slope, whereas the word-final -n curve stands out from the rest. The lexical coverage measure thus nicely illustrates how an ASR tool can highlight linguistic phenomena that involve specific phonemes such as word-final -n in Luxembourgish MND.

4. Summary and prospects

In the present paper, we have highlighted the complex linguistic situation of Luxembourgish, a partially under-resourced and under-described language. We have focused on variants that are elicited by Luxembourgish mobile n-deletion (MND). According to the rule that underlies MND,

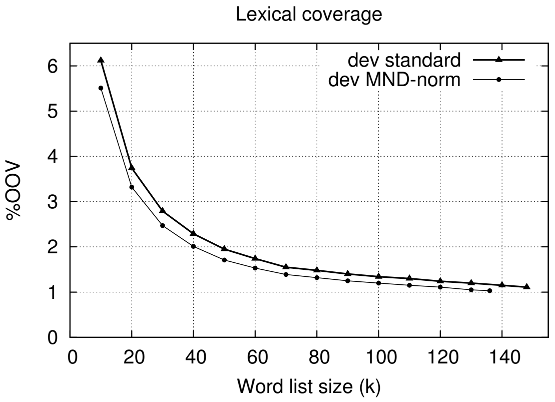

Figure 1: Out of Vocabulary (OOV) word rates measured as a function of word list sizes from the Chamber standard development data and after MND filtering.

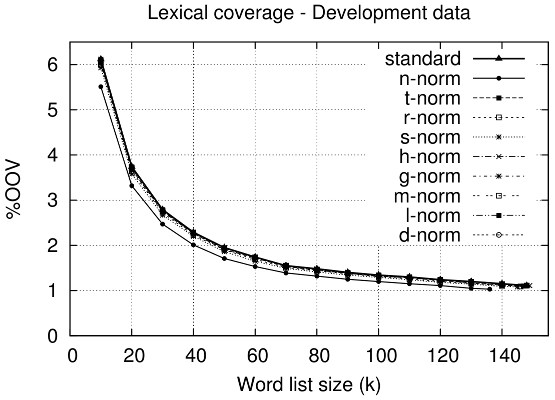

Figure 2: Out of Vocabulary (OOV) word rates measured as a function of word list sizes the from Chamber standard development data and after filtering of various word-final consonants.

word-final -n should be deleted in specific phonological contexts. Thus, MND elicits variants of the same lexical item. Although there are relatively few written resources in Luxembourgish as compared to other languages such as English and German, corpus studies in Luxembourgish will substantially add to the current debate on the processing of variants in automatic and natural speech processing. An important question that is raised by the ASR community, is to know whether the variation is modeled at the lexical level or handled by the acoustic models. It has previously been shown that better recognition performances can be obtained when taking into account the probabilities of pronunciation variants, either at the lexical level or in the acoustic models (Strik, 2001). This information can be readily derived from the type of large corpus-based analyses we are proposing here. Moreover, in order to assess pronunciation and their derived writing variants, it seems that representative data are needed. New methods that are based on pronunciation rules, rather than on the variants directly, can be used to generalize over variants unseen in the training data. From this respect, mobile n-deletion in Luxembourgish provides an excellent test-case, as the variants elicited by MND oc-

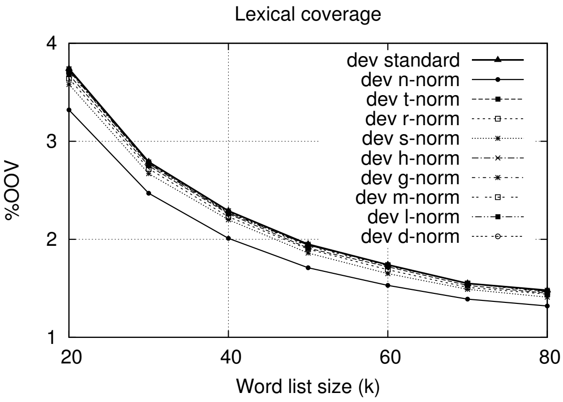

Figure 3: OOV word rates, zoom on the 20-80k word list size range from Figure 2 .

cur in specific phonological contexts and are governed by a linguistic rule. Computational ASR investigations and corpus-based analyses will not only enhance the development of a more full-fledged ASR system for Luxembourgish, but can also be used to highlight specific language phenomena that can make important contributions to linguistic enquiries. For instance, recent research conducted in our lab (Adda-Decker et al., 2010) has looked into the acoustic properties of Luxembourgish by comparing acoustic seed models for Luxembourgish with monolingual German, French, and English acoustic model sets. It was found that German acoustic models provided the best match with the Luxembourgish acoustic models, thereby underpinning the strong Germanic typology of Luxembourgish.

Another important issue pertains to the question as to how listeners cope with pronunciation variants. Indeed, over the last decade a number of studies has looked into perceptual processing mechanisms of variants in spoken word recognition, most notably assimilation of place of articulation (Gaskell and Marslen-Wilson, 1996; Snoeren et al., 2009). Corpus-based studies on variants such as the ones elicited by MND are bound to generate predictions about the representation in the mental lexicon and processing mechanisms that can be readily tested in psycholinguistic experiments. For instance, a critical aspect in the debate on lexical representation and their phonological structure is whether the capacity of distinguishing variants (e.g., those elicited by n-deletion) has to do with auditory perceptual abilities or whether explicit information, i.e. information about the written forms, over the contrastive sounds may be needed to build separate lexical representations. Given the numerous implications and applications that follow from large corpus-based studies, it is hoped that this line of research on Luxembourgish will sparkle more interest for the language in researchers working in the domains of ASR, cognitive psychology, and linguistics.

Acknowledgements

The research has been supported by a grant from the Luxembourgish F.N.R (*Fonds National de la Recherche*) awarded to N.D.S. and has been partially financed by OSEO under the Quaero program.

5. References

G. Adda, M. Adda-Decker, J.L. Gauvain, and L. Lamel. 1997. Text normalization and speech recognition in french. In *Proceedings of the European Conference on Speech Technology, EuroSpeech*.

M. Adda-Decker and L. Lamel. 1999. Pronunciation variants across systems, languages and speaking style. *Speech Communication*, 29:83–98.

M. Adda-Decker, T. Pellegrini, E. Bilinski, and G. Adda. 2008. Developments of letzebuergesch resources for automatic speech processing and linguistic studies. In *LREC*.

M. Adda-Decker, L. Lamel, and N.D.Snoeren. 2010. Initializing acoustic phone models of under-resourced languages: A case-study of luxembourgish. In *SLTU*.

C. Berg and C. Weis. 2005. Sociologie de l'enseignement des langues dans un environnement multilingue. rapport national en vue de l'élaboration du profil des politiques linguistiques éducatives luxembourgeoises. Technical report.

M.G. Gaskell and W.D. Marslen-Wilson. 1996. Phonological variation and lexical access. *Journal of Experimental Psychology: Human Perception & Performance*, 22:144–158.

P. Gilles. 2005. *Phonologie der n-Tilgung im Moselfränkischen ('Eifler Regel'): Ein Beitrag sur dialektologischen Prosodieforschung. Perspektiven einer linguistischen Luxemburgistik - Studien zu Diachronie und Synchronie*. Universitätsverlag WINTER Heidelberg.

S. Greenberg. 1999. Speaking in shorthand - a syllabic-centric perspective for understanding pronunciation variation. *Speech Communication*, 2(4):159–176.

E. Kaisse. 1985. *Connected Speech: The Interaction of Syntax and Phonology*. Academic Press, Orlando.

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.

P. Linden. 1950. *Luxemburger Wörterbuch*. P. Linden, Hofbuchdrucker.

J. Lulling and F. Schanen. 2009. *Luxdico (Lëtzebuergesch/Franséisch, Français/Luxembourgeois*. Edition Shortgen.

E. Riley and A. Ljolje, 1995. *Automatic generation of detailed pronunciation lexicons*, chapter 12. Kluwer Academic Press, Boston.

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.

N.D. Snoeren, M.G. Gaskell, and A.M. Di Betta. 2009. The perception of assimilation in newly learned novel words. *Journal of Experimental Psychology: Learning, Memory & Cognition*, 2(4):542–549.

H. Strik and C. Cucchiarini. 1999. Modeling pronunciation variation for asr: A survey of the literature. *Speech Communication*, 29:115–246.

H. Strik. 2001. Pronunication adaptation at the lexical level. In *ISCA Tutorial and Research Workshop*, Sophia-Antipolis, France.