# Automatic language identity tagging on word and sentence-level in multilingual text sources: a case-study on Luxembourgish

**Thomas Lavergne**[1], **Gilles Adda**[2,4], **Martine Adda-Decker**[2,3], **Lori Lamel**[2]
[1] ILES group, LIMSI-CNRS,
B.P. 133, 91403 Orsay cedex, France, Thomas.Laverge@limsi.fr
[2] Spoken Language Processing group, LIMSI-CNRS,
B.P. 133, 91403 Orsay cedex, France, Gilles.Adda@limsi.fr
[3] Laboratoire de Phonétique et Phonologie
19 rue des Bernardins 75005 Paris, martine.adda-decker@univ-paris3.fr
[4] IMMI-CNRS, B.P. 133, 91403 Orsay cedex, France, adda@immi-labs.org

**Abstract**

Luxembourgish, embedded in a multilingual context on the divide between Romance and Germanic cultures, remains one of Europe's under-described languages. This is due to the fact that the written production remains relatively low, and linguistic knowledge and resources, such as lexica and pronunciation dictionaries, are sparse. The speakers or writers will frequently switch between Luxembourgish, German, and French, on a per-sentence basis, as well as on a sub-sentence level. In order to build resources like lexicons, and especially pronunciation lexicons, or language models needed for natural language processing tasks such as automatic speech recognition, language used in text corpora should be identified. In this paper, we present the design of a manually annotated corpus of mixed language sentences as well as the tools used to select these sentences. This corpus of difficult sentences was used to test a word-based language identification system. This language identification system was used to select textual data extracted from the web, in order to build a lexicon and language models. This lexicon and language model were used in an Automatic Speech Recognition system for the Luxembourgish language which obtain a 25% WER on the Quaero development data.

**Keywords:** under-resourced language; language identification; corpus of Luxembourguish

## 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 population (both natives or residents) generally speak 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 et al., 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 learnt, followed by French.

A consequence of this is that speakers will frequently switch between Luxembourgish, German, and French, on a per-sentence basis, as well as on a sub-sentence level. As an example, a plurilingual sentence such as `Dat hu mier par main levée ofgestëmmt` (This has been voted by a show of hands) mix Luxembourgish and French. In particular, a word list including entries of different languages is problematic when addressing grapheme-to-phoneme conversion as these rules are mostly language dependant.

In order to build resources like lexicons, and especially pronunciation lexicons, or language models needed for natural language processing tasks such as automatic speech recognition, language used in text corpora should be identified. And even if a word, coming from another language, should be considered as a Luxembourgish word (for instance "merci"), the origin of the word will help to build the pronunciation dictionary.

In this paper we present the design of a manually annotated corpus of mixed language sentences as well as the tools used to select these sentences. This corpus of difficult sentences was used to test a word-based language identification system.

We present language identification results at the sentence and sub-sentence levels. Both the corpus[1] and the tools are made freely available to the community.

This language identification system was used to select textual data extracted from the web, in order to build a lexicon and language models. This lexicon and language model were used in an Automatic Speech Recognition system in Luxemburguish which obtain a 25% WER on the Quaero [2] development data (Adda-Decker et al., 2014)

## 2. Motivation

When a large proportion of textual data contains more than one language, filtering them helps the construction of lexicon and language model. A rule-based process can be used as described in (Adda-Decker et al., 2008) but this process has some flaws: it is language and source dependent, and

[1] LuxId corpus, freely available under Creative Commons (CC BY-SA 3.0 FR) licence through the " Share your LRs!" initiative.

[2] `http://www.quaero.org/`

seems to be inefficient in case of heterogeneous multilingual texts such as the ones we find on the web (see for instance Figure 1 which shows a typical web page from a news magazine). Yet, the lexicon and the language model will benefit from the use of such data. Sub-sentence identification is even more difficult to perform with a rule-based system as the number of rules will grow and should take into account lexical entries that can be shared by multiple languages. For short words, combinatorics are reduced and hence short words are often shared across languages without any etymological link: `ville` means "city" /vIl/ in French , and "many" /fIlə/ in Luxembourgish, `net` means "clear,tidy" /nɛt/ in French, and stands for the negation "not" /nœt/ in Luxembourgish. Among the longer words, shared entries generally imply shared origins and semantics. Here one typically finds French or German imports and proper names `Stagiaire, Quartier, Porto, Dubrovnik, Notre-Dame....`
To solve this problem, we decided to use a stochastic language identification system to be able to efficiently filter Luxembourgish from the German, French and English languages.

## 3. Sentence level language identification

An automatic language identification module based on a log-linear maximum entropy (Maxent (Berger et al., 1996)) approach has been used to decide of the language identity on a sentence by sentence basis. As an example, a plurilingual sentence such as `Dat hu mier par main levée ofgestëmmt` (This has been voted by a show of hands) – a typical sentence in Luxembourgish Chamber debates – may be identified as Luxembourgish or rejected as French.
The formulation of the problem is:

$$p(l \mid x) = \frac{\exp\left(\sum_k \theta_k f_k(x, l)\right)}{Z_\theta(x)}$$

where, $l$ is the language, $x$ is the segment (here a sentence), $f_k$ are the features, $\theta_k$ the associated weights, and $Z_\theta(x)$ the partition function. This model is simple and efficient to train, and we could use various interdependent features. The features used in the present experiments are n-grams of characters (sequences of $n$ chars), with $n \in [1, 4]$. In the future, we will add some lexical features as well as some more context information. We evaluated the system on a set of 20k sentences per language, extracted from the WMT News Commentary corpus,[3] with 5 different languages (French, English, German, Spanish and Czech). The results are summarized in Table 1, and exhibit some excellent identification and detection results.

## 4. Manually annotated corpus

We based our method on a corpus coming from the CHAMBER (House of Parliament) debate reports[4] accounting

| #train | Fre,Eng | +Ger | +Spa | +Cze |
|---|---|---|---|---|
| 10k | 3.3‰ | 7.5‰ | 21.1‰ | 21.2‰ |
| 50k | 0.8‰ | 2.4‰ | 9.8‰ | 9.8‰ |
| 100k | 0.3‰ | 1.6‰ | 3.7‰ | 4.2‰ |

| | Fre | Eng | Ger | Spa | Cze |
|---|---|---|---|---|---|
| recall | 0.3‰ | 0.5‰ | 0.1‰ | 0.3‰ | 0.0‰ |

Table 1: (top) language identification errors as a function of training size (in sentences) and number of languages to identify; (bottom) detection of one language (with 100% precision) among the other languages.

| | |
|---|---|
| # of sentences | 924 |
| # of segments | 1510 |
| # of tokens | 8604 |

| | |
|---|---|
| Lux | 825 |
| Fre | 309 |
| Ger | 29 |
| Lux + Fre | 297 |
| Lux + Ger | 47 |
| Lux + Fre + Ger | 3 |

Table 3: Contents of the mixed language corpus, annotated at the segment level

twenty-two millions words; the debates contains some texts in French language (about 25%), some transcriptions from Luxembourgish speech, and a few percent of sentences containing both French and Luxembourgish words. This corpus contains mainly French and Luxembourgish languages, but other Luxembourgish corpus (for instance Web corpus) will contain a mix of Luxembourgish and German languages, or Luxembourgish and Portuguese languages, and so on. Using the Melis tool described in the previous section on the Chamber corpora, we selected 925 sentences where the model was uncertain about the language indicating a sentence with mixed language.
These sentences have been manually split into segments according to the language used, each segment containing a mean of 5.7 tokens. Some segments are annotated with a set of possible language due to the possible sharing of words across languages as illustrated by the samples in Table 2. A summary of the resulting corpus is given in Table 3.

## 5. Word level language identification

For word level language identification we used a linear-chain conditional random field (Lafferty et al., 2001) setup to predict the language of each word taking account of its surrounding context. As the maxent model presented previously, a Conditional Random Field (CRF) is a log-linear model but take account of markovian dependency between consecutive labels:

$$p(l \mid x) = \frac{\exp\left(\sum_t \sum_k \theta_k f_k(x, l_{t-1}, l_t)\right)}{Z_\theta(x)}$$

where $x$ and $l$ are now respectively a sequence of wrods and of language labels.
As some parts of the sentence are ambiguous and may be in more than one language, special care is needed. We choose to model this ambiguity in two ways:

[3] The WMT News Commentary parallel corpus contains news text and commentaries from the Project Syndicate and is provided as training data for the series of WMT translation shared tasks (See http://statmt.org/).

[4] `http://www.chd.lu/wps/portal/public/CompteRenduDesSeances`

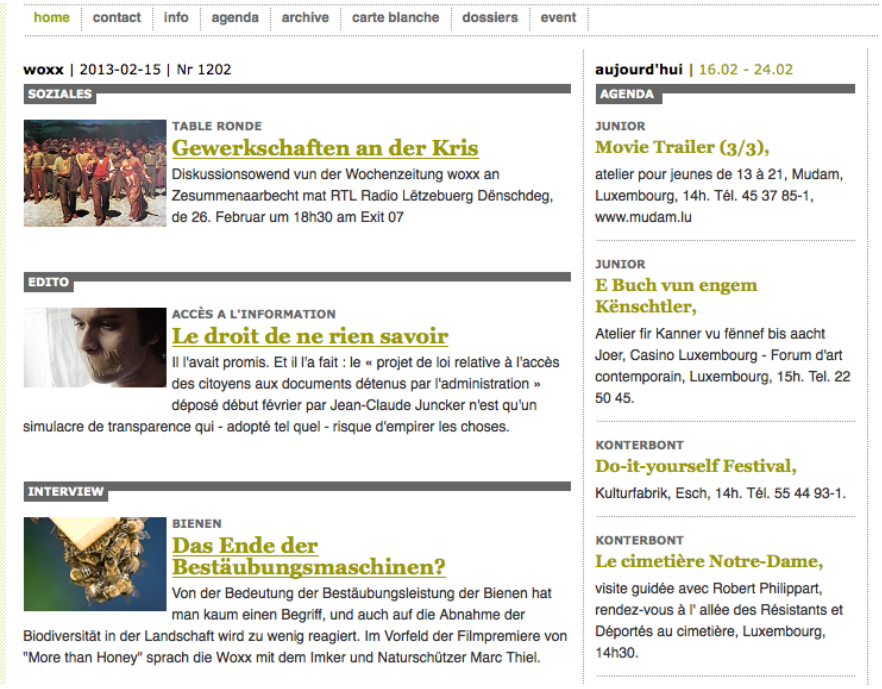


Figure 1: A typical web page from a Luxembourgish news magazine. Source: weekly issue **woxx** `http://www.woxx.lu/`

| |
|---|
| [*LF:* Merci,] [*L:* Här] [*LG:* Minister] |
| [*L:* De] [*LG:* Respekt] [*L:* vun] [*LG:* der Regierung] [*LF:* par rapport] [*LG:* zum Parlament] [*L:* gebitt dat.] |
| [*L:* Ech bieden Iech,...] [*LF:* M. Jean-Marie Halsdorf,] [*F:* Ministre de l'Intérieur et à la Grande Région.] |
| [*L:* Parlamentaresch] [*LFG:* Versammlung] [*L:* vum Europarot] |
| [*L:* Ech hale fest, dass d'Vertrieder vum] [*F:* Ministère de la Famille] [*L:* net iwwerzeegt dovu waren, dass dat néideg wär. Véiertens, et soll een dem] [*F:* Office social] [*L:* de] [*LFG:* Statut] [*F:* d'établissement public communal sous la surveillance de la commune et le contrôle de l'État ginn.] |

Table 2: Sample sentences from the corpus showing the different kind of ambiguity.

- Using special labels representing the different combinations of languages for a total of six different labels. In this case, the CRF is a classical linear-chain CRF and can be trained as usual.
- using only three labels but allowing more than one of them to be valid in the reference. In this case, the training is a little more involved as the computation of the empirical expectation should take account of the multiple reference labelling.

The evaluation is done using ten-fold cross-validation due to the small size of the corpus.
We trained to CRFs using the Wapiti toolkit(Lavergne et al., 2010) with two different set of features. The *baseline* contains unigrams and bigrams features of words in a context window of size 5. The *presuf* add prefixes and suffixes of words upto to 4 characters.
Results are summarized in Table 4. Two different error rates are reported:

- (A) predictions are considered good only if the system has predicted exactly the set of languages associated with the word;

| System | Err (A) | Err (B) |
|---|---|---|
| 6-labels: *baseline* | 12.3% | 9.9% |
| 6-labels: *presuf* | 10.1% | 7.6% |
| 3-labels: *baseline* | | 9.0% |
| 3-labels: *presuf* | | 7.1% |

Table 4: Language identification error rates on a word level using the CRF. (A) stands for exact L-set identification; (B) stands for L-subset identification

- (B) predictions are considered good if the system has predicted a correct subset of languages associated with the word.

For the 3-labels setup, only the (B) scores are reported as this system may only precict one language at each positions.

## 6. Automatic speech recognition in Luxembourgish

First results of large vocabulary continuous speech recognition (LVCSR) for Luxembourgish were presented in (Adda-

Decker et al., 2011) on set of manually transcribed data (70 minutes from CHAMBER and 10 minutes from RTL). The word error rates (WER) were in the range of 55 to 70%. In order to obtain recognition word error rates close to those reported for other European languages, it is necessary to estimate the acoustic models on substantially more audio data. Unfortunately, however, no speech corpora with manual transcripts are available for Luxembourgish. Therefore it was decided to apply the semi-supervised acoustic model training developed in (Lamel et al., 2002b; Lamel et al., 2002a). The basic idea is to iteratively automatically transcribe a large volume of Luxembourgish speech data, providing indirect supervision via the language model. A detailed description could be find in (Adda-Decker et al., 2014)

We collected different Luxembourgish texts, some described in (Adda-Decker et al., 2008) and others newly collected from the web. The texts belong to 3 domains:

1. **'New/information' related written sources**:
   - RTL2008: old RTL data (2008 and earlier) manually filtered.
   - RTL2012: Web sites affiliated to RTL (collected in 2012).
   - WIKIPEDIA: Luxembourgish Wikipedia.
   - MISC: miscellaneous reports, books, reviews ... collected on the web.
2. **Oral transcriptions**:
   - CHAMBER: *bona fide* transcriptions (Adda-Decker et al., 2008) of the Luxembourgish Parliament debates.
3. **Social media**:
   - BLOGS: 90 blogs (out of 400 preselected Luxembourgish blogs).
   - BLOGS_COMMENT: user comments from the selected blogs.

Thez language identification system described above was used to efficiently filter Luxembourgish texts from those in the German, French and English languages in order to process heterogeneous multilingual texts such as are typically harvested from the Web.

The volume of raw texts and of filtered texts are summarized in Table 5. The amount of rejected data (average 33%) strongly depends on the source as expected: for WIKIPEDIA only 3% of the data were rejected[5], while 68% of the Luxembourgish BLOGS were not written in Luxembourgish, according to the automatic identification system, even though only the blogs (90 out of 400) with a significant part of written Luxembourgish were kept. 27% of the CHAMBER texts were also rejected: beyond transcripts in French language due to occasional switches to French language in oral debates, this rather high rejection rate is due to the presence of reports written in French. After filtering, the amount of Luxembourgish-labeled data sums to over 34 Mwords, with an average rejection rate of 33% of the raw texts.

| source | size | size | %rejected |
|---|---|---|---|
| RTL2008 | 611 | 607 | <1 |
| WIKIPEDIA | 3603 | 3483 | 3 |
| RTL2012 | 10,307 | 7948 | 23 |
| BLOGS_COMMENTS | 3106 | 2386 | 23 |
| CHAMBER | 22,110 | 16,108 | 27 |
| MISC | 1677 | 855 | 49 |
| BLOGS | 10,243 | 3265 | 68 |
| total | 51,657 | 34,653 | 33 |

Table 5: Text size (in thousands of words). (left) Raw texts per data source (7 sources, totaling 51Mwords) (right) Luxembourgish text sizes and percentages of rejected texts.

Both the raw and filtered texts were used to build and compare word lists and language models, using the methods described in (Adda-Decker et al., 2008). The 200k most probable words were selected from the 7 Web data sources, so as to minimize the unigram perplexity. An OOV (Out of Vocabulary) rate of 2.35% was achieved with the filtered sources, to be compared to an OOV rate of 3.23% with the raw texts (28% relative improvement). With respect to the language model, the best interpolated 3-gram model gives a dev set perplexity of 369.35 with the filtered sources (387.20 without filtering, +5%). Due to filtering, the OOV rate exhibits a large improvement, with a more limited gain for the language models. This is generally observed when the amount of texts is insufficient: filtering improves the precision of the word list, however the negative impact on perplexity of filtering out few correct Luxembourgish n-grams counterbalances the positive impact of improving the precision.

Using the filtered language model, 1200 hours of untranscribed audio data were used to train acoustic models in an iterative manner, progressively increasing the quantity of audio. Using these acoustic models and the filtered language model, a 25.6% WER on the Quaero 2013 development data has been obtained (Adda-Decker et al., 2014).

## 7. Conclusion

We present the development of a sub-sentence language identification system for Luxembourgish. Texts in Luxembourgish language present frequent switches between Luxembourgish and another major languages (usually French or German). The development of efficient models for natural language processing tools in Luxembourgish requires taking care of this phenomenon.

Our approach is firstly to identify the language of the whole sentence, in order to select the ambiguous ones. Next, we label each of the words of these ambiguous sentences with a set of possible languages. The Maxent model used for the first step exhibits very good results. The second step is handled by a linear-chain CRF which directly models the language ambiguity to improve performances over a baseline CRF. Future work on this step will include more complex features sets like including lexicon for each language.

[5] some residual non-Luxembourgish languages such as ancient Greek was rejected because of its special coding alphabet

This language identification system was used to select textual data extracted from the web, in order to build lexicon and language models. This lexicon and language model were used in an Automatic Speech Recognition system in Luxemburguish which obtain a 25% WER on the Quaero development data (Adda-Decker et al., 2014).
In the process of developing this framework, a corpus of 924 Luxembourgish sentences manually annotated at the word level has been made freely available.

## 8. Acknowledgments

This work has been partially financed by OSEO under the QUAERO program, and supported by the French Investissements d'Avenir - Labex EFL program (ANR-10-LABX-0083).

## 9. References

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

Adda-Decker, M., Lamel, L., and Adda, G. (2011). A first LVCSR system for Luxembourgish, an under-resourced European language. In *LTC'11 LTC-LRL workshop, 5th Language & Technology Conference*, Poznan, Poland.

Adda-Decker, M., Lamel, L., and Adda, G. (2014). Speech alignment and recognition experiments for luxembourgish. In *Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU)*, St Petersburg, Russia.

Berger, A. L., Pietra, V. J. D., and Pietra, S. A. D. (1996). A maximum entropy approach to natural language processing. *Comput. Linguist.*, 22(1):39–71.

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.

Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proceedings of the International Conference on Machine Learning*, pages 282–289. Morgan Kaufmann, San Francisco, CA.

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

Lamel, L., Gauvain, J.-L., and Adda, G. (2002b). Unsupervised acoustic model training. In *ICASSP*, pages 877–880.

Lavergne, T., Cappé, O., and Yvon, F. (2010). Practical very large scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 504–513. Association for Computational Linguistics, July.

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