A Limburgish Corpus Dictionary: Digital Solutions for the Lexicography of a Non-standardized Regional Language

Yuri Michielsen-Tallman[1], Ligeia Lugli[2], Michael Schuler[3]

[1]Maastricht University, FASoS, Grote Gracht 90, 6211 SZ Maastricht.

[2]King's College London, Virginia Woolf Building, 22 Kingsway, London, WC2B 6LE.

[3]Unaffiliated, now at Google Inc.

E-mail: j.michielsen@maastrichtuniversity.nl, ligeia.lugli@kcl.ac.uk, masmas@google.com

Abstract

This paper presents the Limburgish Corpus Dictionary (LCD), a newly-started project at Maastricht University that aims to create an online corpus and dictionary of Limburgish from scratch.

Limburgish comprises a set of West Germanic dialects spoken in the Dutch and Belgian provinces of Limburg. Due to a variety of factors, including its history and geographic spread, Limburgish exhibits an extremely high degree of spelling variation. In conformity with current policies, our dictionary strives to give equal visibility to all local dialects and variant spellings, with a view to enabling users to search for and retrieve lexical entries using their preferred spelling of a lemma.

After a brief outline of the Limburgish language, the history of writing in Limburgish, and Limburgish lexicography, this paper presents the dynamic and multi-layered entry structure that we have devised to represent information about spelling variation. Subsequently, it discusses how our lexicographic model impacts the way we prepare our corpus for analysis. It concludes with a description of our tentative corpus-processing pipeline and the results of some initial NLP software testing.

Keywords: minority language; Limburgish; spelling variation; normalization; lemmatization

1. Introduction

This paper introduces the Limburgish Corpus Dictionary (LCD), a project recently started at Maastricht University in cooperation with the *Meertens Instituut* and other partners. Much befitting the eLex theme of this year, this project starts completely from scratch. Despite a long history of Limburgish lexicography, the LCD will be the first lexicographic resource of its kind. It is the first dictionary to be derived from a digitized corpus of texts written in Limburgish and the first to include all spelling variations found in varieties of Limburgish. This requires unprecedented efforts and raises new challenges. In this paper, we focus only on those efforts and challenges that stem from the lack of an agreed upon standard written variety and the consequent abundance of co-existing spelling variants for every lemma.

The paper comprises four parts. First, it opens with a brief overview of Limburgish, its writing and spelling practices, and lexicography history. It proceeds to describe a model to represent different dialectal varieties in a single online dictionary.

Subsequently, it outlines how spelling variation complicates corpus processing and describes a set of heuristics and computational tools available to address these issues. Finally, it delves into future lines of development, especially regarding a possible NLP software pipeline.

2. Limburgish

2.1 Limburgish language

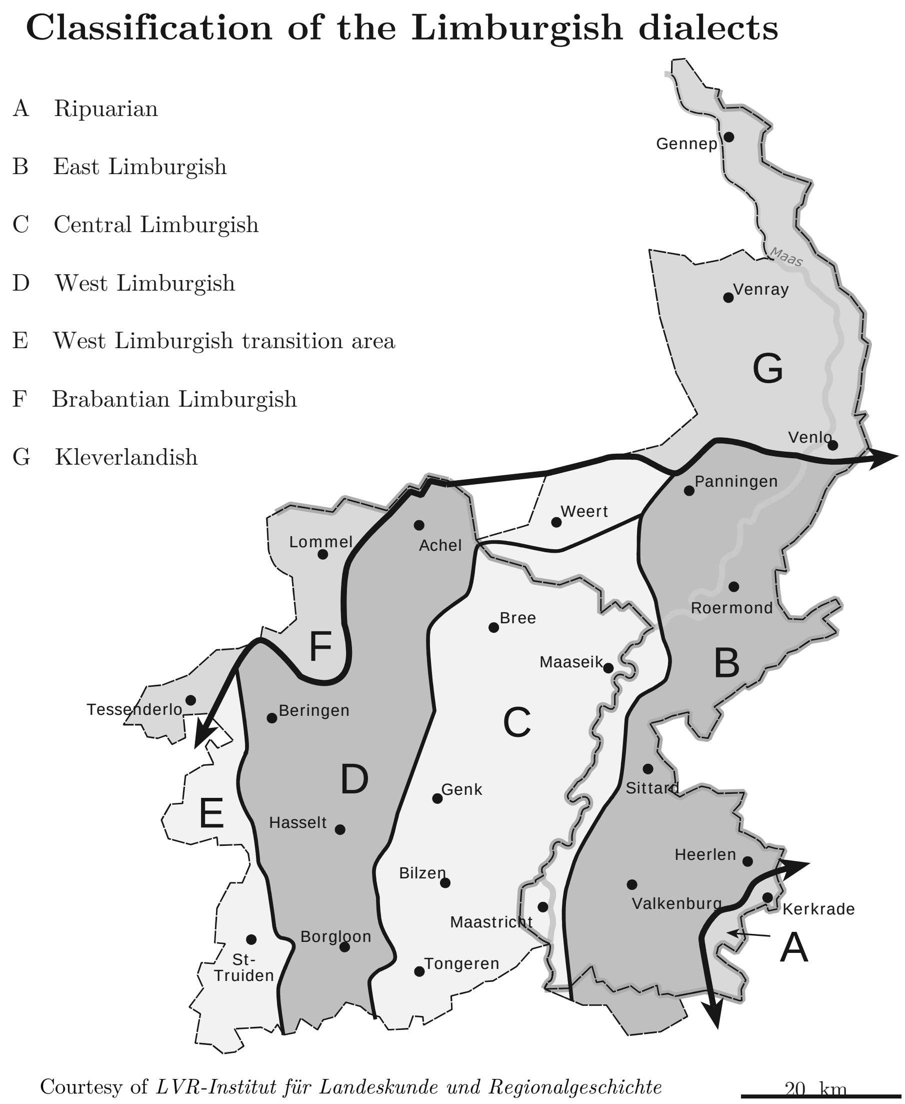
Limburgish refers to a language variety that is part of a continuum of West Germanic dialects, traditionally referred to as East Low Franconian in Dutch and Flemish dialectology and South Low Franconian in German dialectology (Belemans, 2009: 29). Limburgish consists of several dialects that share fundamental common characteristics (Schutter & Hermans, 2013), are mutually intelligible (Leerssen et al., 1996), and exhibit linguistic variety (Draye, 2007: 15). Its demarcation is subject to debate, but in many definitions Limburgish refers to most, though not all, of the dialects spoken in the Dutch and Belgian provinces of Limburg and some adjacent areas in the German Rhineland region, delimited by the Ürdinger isogloss (*ik-ich*) and the Benrather isogloss (*maken-machen*) (Belemans, 2009: 14; Notten, 1988: 71). For the purposes of the Limburgish Corpus Dictionary (LCD) we will adhere to the demarcation of Limburgish as used by the *Woordenboek van de Limburgse Dialecten* (Dictionary of the Limburgish dialects)[1] and illustrated below in Figure 1.

Limburgish developed separately from other Low Franconian varieties. It has a different phonetic system, grammar, and vocabulary. Unlike other Low Franconian varieties it only marginally contributed to the development of standard Dutch (Opgenort, 2012; Leerssen et al., 1996). According to some measures, the dialects of Limburgish are further removed from standard Dutch than any dialect or other regional language in the Netherlands and the Dutch-speaking part of Belgium (Hoppenbrouwers & Hoppenbrouwers, 2001; van Hout & Münstermann, 1981). Moreover, strikingly different from Dutch, as part of a continuum of Low and Central Franconian tonal dialects, most Limburgish dialects exhibit binary tone contrast on long vowels and diphthongs (Boersma, 2013; Gussenhoven & Peters, 2008; Fournier et al., 2004).

In the Netherlands, since 1997, Limburgish has enjoyed some official recognition as a regional language according to Part II European Charter for Regional or Minority Languages (Swanenberg, 2013). This legal recognition applies to all dialects spoken in the province of Dutch Limburg. This includes the small Kleverland and Ripuarian dialect regions that under some definitions are viewed as part of respectively Brabantian-Dutch and High German dialects (see below Figure 1) (Belemans et al., 1998; Daan & Blok, 1969). As part of this recognition, at the regional level, the

[1] See Belemans et al. (1998) and Weijnen et al. (1983: 7-11, 22).

Dutch province of Limburg has established an advisory body *Raod veur 't Limburgs* (Council for Limburgish) to tend to Limburgish. However, this is not the case in Belgium and Germany, where Limburgish has no official status.



Courtesy of *LVR-Institut für Landeskunde und Regionalgeschichte*

Figure 1: Map of the main Limburgish dialect areas

2.2 Written Limburgish

Since the LCD is a based on a diachronic written corpus (see 2.4 below), a brief history of writing in Limburgish against the backdrop of Limburg's history might be useful. Writing in Limburgish has a long history. The Wachtendonck Codex of around 900 CE contains the oldest known Limburgish fragment (Jongen, 2016: 25; Robinson, 1992: 205). During the Middle Ages, Limburgish was an important literary language (Tervooren, 2006) and was used as a language of government and administration (Willemyns, 2003; Moors, 1952). Wars fought in the territories of present-day Limburg during the 16th and 17th centuries led to increasing political fragmentation, due to which either French, German or Dutch replaced Limburgish as a language of government (limburgs.org; Otten, 1977). As a result of economic and cultural decline, literary production stagnated (van Horen & van Horen-Verhoosel, 2016: 67). In 1795 the fragmented Limburgish territories were unified and incorporated by France as a *département*. Subsequently, in 1815, they were placed under Dutch control by the Congress of Vienna. During the Belgian uprising in 1830, Limburg seceded to become part of Belgium. In 1839, the east of Limburg was returned to the Netherlands, splitting the region into a Dutch and a Belgian province. For reasons that are unclear, at the end of the 18th century, writing in Limburgish slowly revived (Spronck, 1962: 436). From 1840 onwards, literary production started gathering pace (Spronck, 2016; Nissen, 1986), especially in literary societies in the urban centers of Dutch Limburg. In 1926 with the foundation of Veldeke, a Limburg-wide organization to promote the use of Limburgish, writing in Limburgish became more common practice (Spronck, 2016).

2.3 Spelling variation

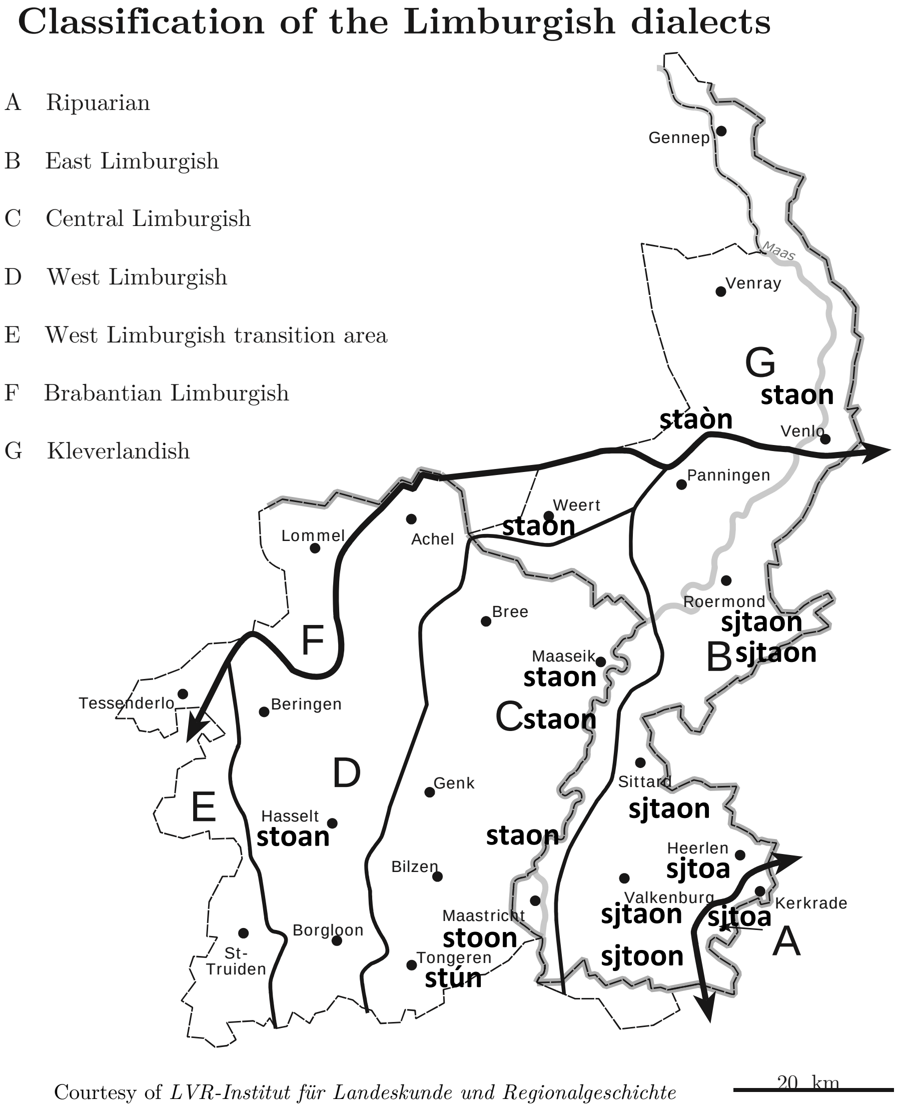
Spelling variation is very much part of Limburgish writing. Possibly as a result of its past political fragmentation, Limburgish speakers strongly identify with their native locality and its dialect. Virtually all published (or online) texts are accompanied by an indication of the dialect that is used. This practice both testifies to and likely reinforces such identification. An attempt to unify the written standard faltered in the Limburgish parliament in 2000 (limburgs.org).

The official policy of the Council for Limburgish is to treat all dialects of Limburgish equally (Weusten et al., 2013; van Hout, 2007) and to support the current variation in spelling practices. To this end, in 2003, the Council for Limburgish created a normative orthography, which links graphemes and phonemes and can be used for writing in the different Limburgish dialects (Opgenort, 2012; Bakkes et al., 2003). This orthography is based on a succession of previous spelling guidelines created by Veldeke, the main regional language organization, since 1934 (Wolters, 2016), which in turn was influenced by the orthographic tradition that developed in the wake of the literary revival of the 19th century. Much, though not all, of the writing since 1934 is based on the Veldeke guidelines (limburgs.org). Yet, this does not ensure spelling homogeneity, and the result is a phonological and sometimes idiosyncratic spelling that reflects each writer's own dialectal pronunciation and spelling practices. An example of some of the regional spelling variation, based on local dictionary forms, is given in Table 1 and illustrated in Figure 2.

Hasselt	Tongeren	**Maastricht**	Weert	Maasbree	Thorn	Elsloo	Echt
stoan	stún	**stoon**	staon	staòn	staon	staon	staon

Venlo	Sittard	Roermond	Posterholt	Valkenburg	Simpelveld	Heerlen	Kerkrade
staon	Sjtaon	sjtaon	sjtaon	sjtaon sjtoon	sjtoa	sjtoa	sjtoa

Table 1: Representation of spelling variation of some Limburgish dialect-specific lemmas associated with the Maastricht lemma <stoon> [stʊ·²n] 'to stand' taken from local dialect dictionaries of Belgian and Dutch Limburg.



Courtesy of *LVR-Institut für Landeskunde und Regionalgeschichte*

Figure 2: Illustration of spelling variation of some Limburgish dialect-specific lemmas associated with the Maastricht lemma <stoon> [stʊ·²n] 'to stand' taken from local dialect dictionaries of Belgian and Dutch Limburg (lemma forms added to original table)

2.4 Limburgish lexicography

Glossaries of Limburgish dialects exist from the Middle Ages (Jongen, 2016: 25). Since the end of the 19th century around 80 dictionaries and glossaries of local dialects of Limburgish have been created. These vary in size and the methodology used, but virtually all are bilingual to or from Dutch. For the Limburgish content, most adhere to the spelling guidelines mentioned above, applied to the local variant. A few are online[2].

So far, only three lexicographic projects have covered all dialects in Limburg; the *Woordenboek van de Limburgse Dialecten,* the *Taal van de Maas*, and the Limburgish Academy dictionaries. The *Woordenboek van de Limburgse Dialecten* (Dictionary of the Limburgish dialects), completed in 2008, is a thematically-organized dictionary created by the universities of Nijmegen and Leuven. Sources for the dictionary were questionnaires, dictionaries of local dialects and other sources that included research focused on the lexicon. The spelling of the Limburgish lexicon is adapted to standard Dutch, whereby the original Limburgish is spelled according to Dutch phonology and orthography. An online version is available[3]. In the 1990s, the *Werkgroup Algemeen Geschreven Limburgs* (working group General Written Limburgish) created the *Taal van de Maas* (Language of the Meuse), a Dutch–Limburgish dictionary (Prikken, 1994). Its sources and the selection criteria for the Limburgish lexicon are unclear. A spelling system was developed that differed from traditional Limburgish spelling in that it was not based on phonology. An online version gives access to Dutch–Limburgish and Limburgish–Dutch word lists[4]. Finally, on the basis of written and online sources, the Limburgish Academy Foundation created two online dictionaries: a Limburgish–Dutch and a Limburgish–English dictionary. The spelling of Limburgish words is mostly based on the 2003 normative orthography of the Council for Limburgish applied to phonology of the Maastricht dialect. These dictionaries are only available online[5].

The LCD will be the first corpus-driven dictionary of Limburgish. It is based on ideally every extant sample of written, transcribed from spoken, internet, and social media text in every dialect from both provinces of Limburg and the Limburgish territories that preceded their existence[6]. The corpus will be diachronic, encompassing texts from about 1775 until the present, though most texts date from 1926 until the present. The LCD will be a free online dictionary. In line with Limburger writing practices and the official position of the Council for Limburgish, the LCD will strive to give equal representation to all dialectal varieties in Dutch Limburg, as well as Belgian Limburg, and the resultant spelling variation. This has some important lexicographical implications.

[2] See for Gronsveld woordenboek.gronsveld.com, Maastricht mestreechtertaol.nl, and Thorn limburgsewoordenboeken.nl.

[3] See e-wld.nl.

[4] See limburghuis.nl.

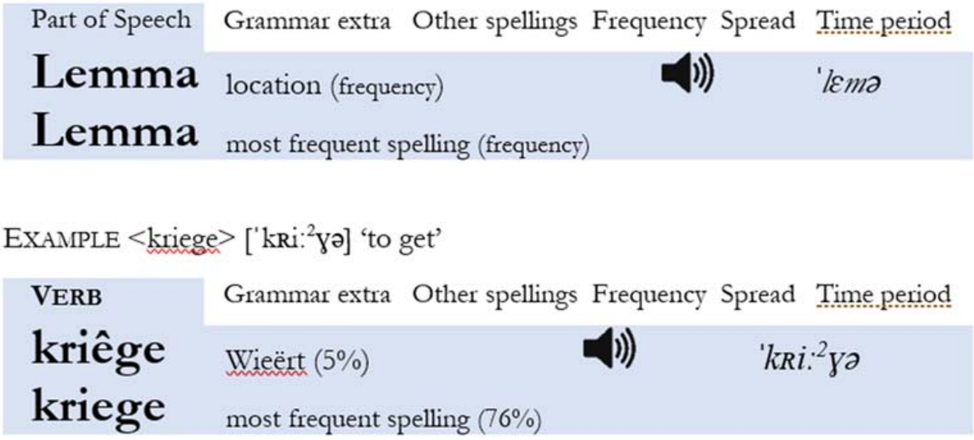
[5] See limburgs.org.

[6] For a complete demarcation of Limburgish we use the definition of the Dictionary of the Limburgish dialects (see 2.1 above).

3. Requirements for a Limburgish Corpus Dictionary

Spelling variation, Limburger writing practices, as well as language policy, all impact our project on the level of the designs of both corpus and dictionary. The lexicographer needs to be able to retrieve all instances of a lemma in the corpus, determine how they are distributed, and identify whether the variation is purely formal or somehow correlates with semantic variation. This calls for processing our corpus in a way that clusters all spelling variation under a single lemma form. The users of our dictionary need to retrieve an entry for a word, regardless of which local spelling they enter in the search box. This would necessitate the possibility of displaying headwords in all the local spelling variations to allow users to see 'their' preferred spelling in the online dictionary.

The LCD is aimed at a range of audiences spanning from general Limburgish-speaking users to linguists. Its primary focus is on non-specialist Limburgish users who will be interested in referencing only limited information in each entry. To facilitate perusal of the dictionary on the part of such non-specialist users, search results will only display the lemma in the user's preferred spelling. In addition to that spelling, the dictionary entry will also display the most frequent spelling of that lemma in the corpus (for problems related to calculating the relative frequency of different spellings of a lemma see Section 4.3 below) to inform the user of a more general spelling of the term throughout Limburgish (see Figure 3).



Part of Speech Grammar extra Other spellings Frequency Spread Time period

Lemma location (frequency) 🔊 'lɛmə

Lemma most frequent spelling (frequency)

EXAMPLE <kriege> [ˈkʀiː²ɣə] 'to get'

VERB Grammar extra Other spellings Frequency Spread Time period

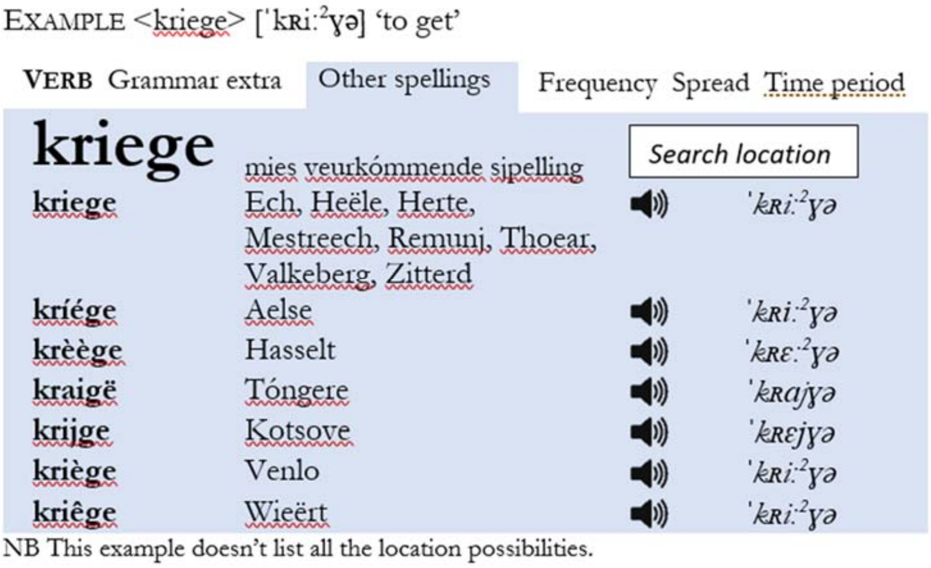
kriêge Wieërt (5%) 🔊 'kʀiː²ɣə

kriege most frequent spelling (76%)

Figure 3: Representation of the display of a lemma in the online dictionary of the user's spelling and the most frequent spelling of that lemma

Users interested in accessing more information about a lemma will be able, by clicking on a tab, to access all spelling variations of a Limburgish lemma as attested by the corpus, including the location[7] where this variant is found and its frequency in the corpus (see Figure 4).

[7] Based on authors' practice in indicating the dialect of a written text, we assign a location with a Kloeke code, a location code commonly used in Netherlandic dialectology: meertens.knaw.nl/kloeke.



EXAMPLE <kriege> ['kʀiː^{2}ɣə] 'to get'

VERB Grammar extra Other spellings Frequency Spread Time period

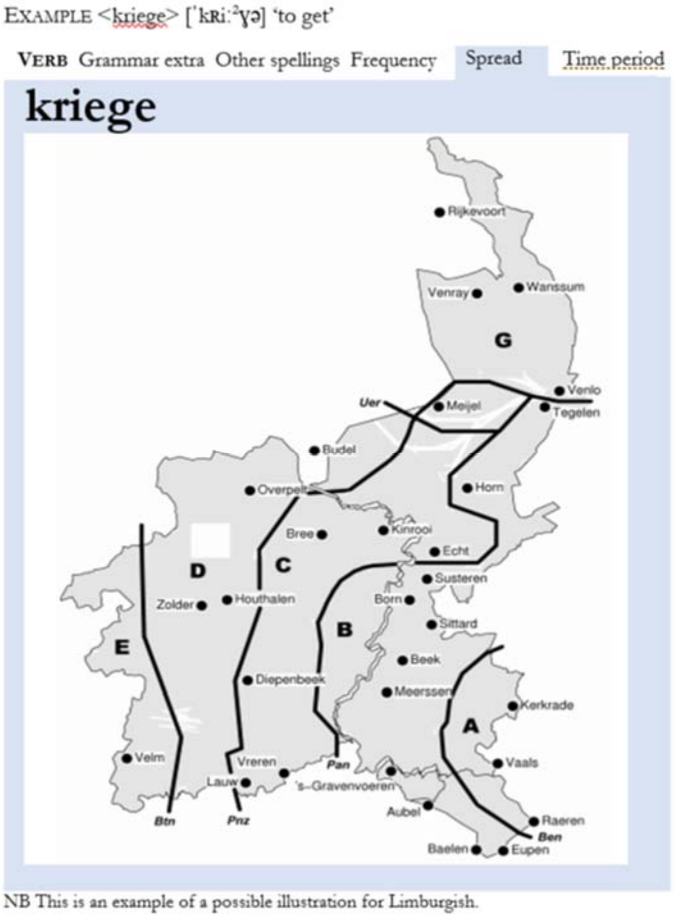
kriege

	mies veurkómmende sjpelling	Search location	
kriege	Ech, Heële, Herte, Mestreech, Remunj, Thoear, Valkeberg, Zitterd	🔊	'kʀiː^{2}ɣə
kríége	Aelse	🔊	'kʀiː^{2}ɣə
krèège	Hasselt	🔊	'kʀɛː^{2}ɣə
kraigë	Tóngere	🔊	'kʀajɣə
krijge	Kotsove	🔊	'kʀɛjɣə
kriège	Venlo	🔊	'kʀiː^{2}ɣə
kriêge	Wieërt	🔊	'kʀiː^{2}ɣə

NB This example doesn't list all the location possibilities.

Figure 4: Representation of the display of spelling variety of a lemma in the online dictionary

Two further viewing modalities will be available to access information about the geographic spread of a lemma throughout Limburg as attested in the corpus (see Figure 5) and a diachronic table indicating the time period of a lemma (see Figure 6).



EXAMPLE <kriege> ['kʀiː^{2}ɣə] 'to get'

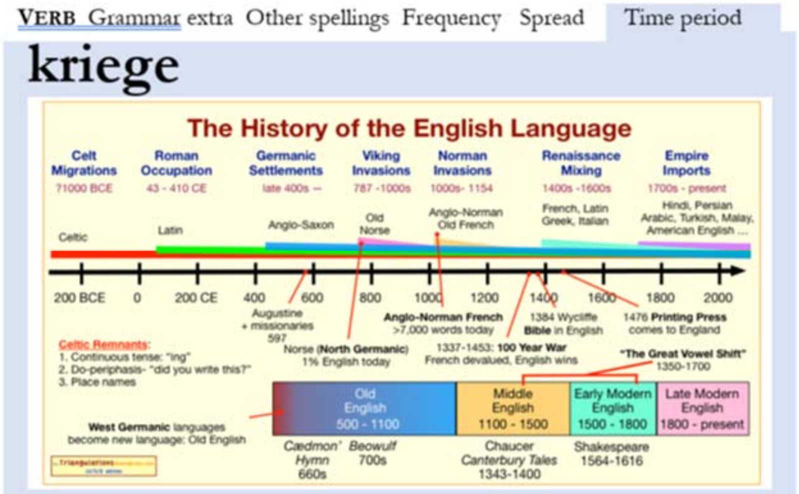
VERB Grammar extra Other spellings Frequency Spread Time period

kriege

NB This is an example of a possible illustration for Limburgish.

Figure 5: Representation of the geographic spread of a lemma in the online dictionary

EXAMPLE <kriege> ['kʀiː²ɣə] 'to get'

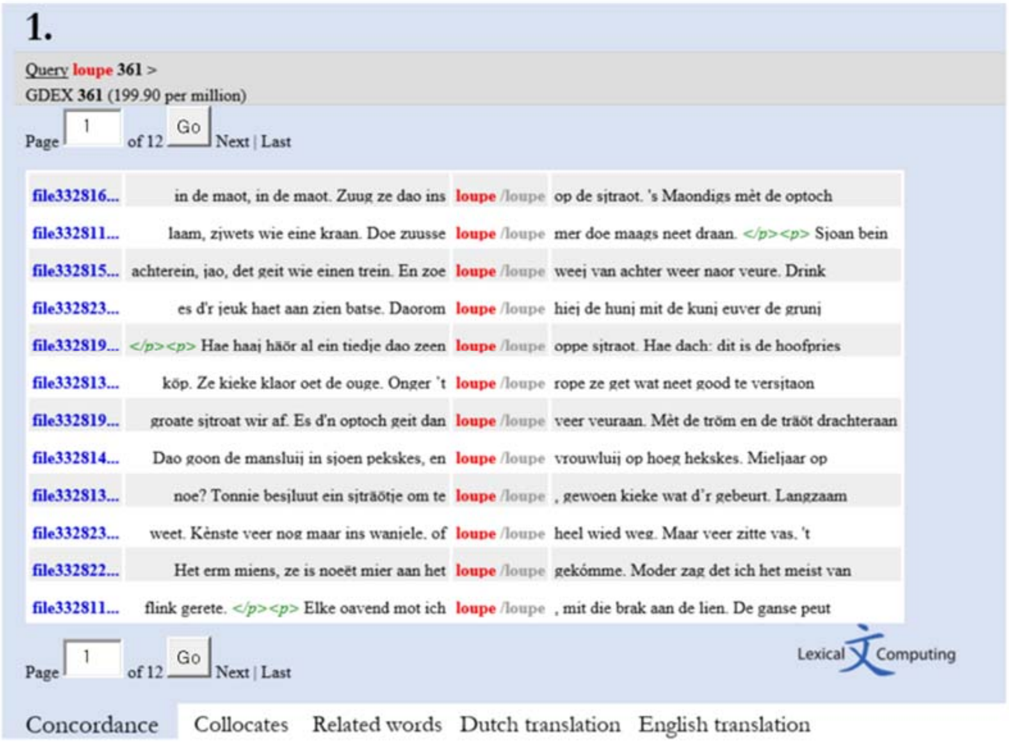

VERB
Grammar extra
Other spellings
Frequency
Spread
Time period
kriege
The History of the English Language
Celt Migrations
?1000 BCE
Roman Occupation
43 - 410 CE
Germanic Settlements
late 400s –
Viking Invasions
787 -1000s
Norman Invasions
1000s- 1154
Renaissance Mixing
1400s -1600s
Empire Imports
1700s - present
Celtic
Latin
Anglo-Saxon
Old Norse
Anglo-Norman Old French
French, Latin Greek, Italian
Hindi, Persian Arabic, Turkish, Malay, American English ...
200 BCE
0
200 CE
400
600
800
1000
1200
1400
1600
1800
2000
Augustine + missionaries 597
Norse (North Germanic) 1% English today
Anglo-Norman French >7,000 words today
1337-1453: 100 Year War French devalued, English wins
1384 Wycliffe Bible in English
1476 Printing Press comes to England
"The Great Vowel Shift" 1350-1700
Celtic Remnants:
1. Continuous tense: "ing"
2. Do-periphasis- "did you write this?"
3. Place names
West Germanic languages become new language: Old English
Old English 500 - 1100
Middle English 1100 - 1500
Early Modern English 1500 - 1800
Late Modern English 1800 - present
Cædmon' Hymn 660s
Beowulf 700s
Chaucer Canterbury Tales 1343-1400
Shakespeare 1564-1616

NB This is an example of a possible illustration for Limburgish (history will be Limburgish-specific).

Figure 6: Representation of a possible display of the time period of a lemma in the online dictionary

Finally, the online dictionary will provide a 'concordance feature' where lexicographers and linguists, after log-in, will have direct access to the corpus (see Figure 7). This feature will be required to portray Limburgish texts in their original spellings.

EXAMPLE CONCORDANCE FEATURE AFTER LOGIN <loupe> ['lɔ·²pə] 'to walk'


1.

Query loupe 361 >
GDEX 361 (199.90 per million)

file332816...	in de maot, in de maot. Zuug ze dao ins	loupe /loupe	op de sjtraot. 's Maondigs mèt de optoch
file332811...	laam, zjwets wie eine kraan. Doe zuusse	loupe /loupe	mer doe maags neet draan. </p><p> Sjoan bein
file332815...	achterein, jao, det geit wie einen trein. En zoe	loupe /loupe	weej van achter weer naor veure. Drink
file332823...	es d'r jeuk haet aan zien batse. Daorom	loupe /loupe	hiej de hunj mit de kunj euver de grunj
file332819...	</p><p> Hae haaj häör al ein tiedje dao zeen	loupe /loupe	oppe sjtraot. Hae dach: dit is de hoofpries
file332813...	köp. Ze kieke klaor oet de ouge. Onger 't	loupe /loupe	rope ze get wat neet good te versjtaon
file332819...	groate sjtroat wir af. Es d'n optoch geit dan	loupe /loupe	veer veuraan. Mèt de tröm en de träöt drachteraan
file332814...	Dao goon de mansluij in sjoen pekskes, en	loupe /loupe	vrouwluij op hoeg hekskes. Mieljaar op
file332813...	noe? Tonnie besjluut ein sjträötje om te	loupe /loupe	, gewoen kieke wat d'r gebeurt. Langzaam
file332823...	weet. Kènste veer nog maar ins wanjele. of	loupe /loupe	heel wied weg. Maar veer zitte vas. 't
file332822...	Het erm miens, ze is noeët mier aan het	loupe /loupe	gekómme. Moder zag det ich het meist van
file332811...	flink gerete. </p><p> Elke oavend mot ich	loupe /loupe	, mit die brak aan de lien. De ganse peut

Figure 7: Representation of the 'concordance feature' for the lemma <loupe> [ˈlɔ·²pə] 'to walk' in the online dictionary

To enable this entry structure in the dictionary and to allow lexicographers to retrieve and analyze all the relevant information about spelling variation, we outline the following considerations to ensure that our corpus is adequately processed.

4. NLP tools and Limburgish spelling variation

4.1 NLP tools and spelling variation

NLP tools have mostly been developed to process standardized languages and are not designed to deal with languages rich in spelling variation. Several NLP tools have been developed to process spelling variation, especially for historical corpora (see, e.g., van Halteren & Rem, 2013). The main pathway has been to apply a preprocessing tool before lemmatizers or Part of Speech (PoS)-taggers to normalize all orthographic variants of a token to a single spelling (Barteld et al., 2016). This normalization leads to more accurate processing in subsequent NLP tools, (Hendrickx & Marquilha, 2011). This practice presumes the existence of a standardized language that can be used for normalization. For standardized languages, unary normalization of diachronic corpora is possible, but has also proven problematic (Archer et al., 2015). For a non-standardized contemporary language like Limburgish, the issues are more complex. We will first outline some general issues pertaining to corpus normalization and lemmatization that have arisen in our project, and we will then describe a tentative processing pipeline and the result of some initial software testing.

4.2 Normalization for spelling variation in Limburgish

Our corpus exhibits both diachronic and synchronic spelling variation. Its diachronic and multi-dialectal nature, combined with idiosyncratic spellings and the lack of an agreed upon written standard, lead to an extremely high degree of spelling variation in a Limburgish corpus. This problem is by no means unique to this project. It has indeed already been treated effectively within several other projects, mostly of a historical nature, where the texts were normalized to a single standardized variety of the language, typically the contemporary form of the language[8].

In our project, however, the policy of treating all dialectal varieties equally adds a layer of complexity to the task of corpus normalization. The rationale for text-normalization is that in other cases it facilitates information retrieval because the language to which the text is normalized is more standardized and more widely accessible than the original. In the case of Limburgish, however, we face a multitude of similarly non-standardized varieties, none of which is more universally accessible than the others.

[8] For a survey of technical approaches used for normalizing historical texts see e.g. Barteldet al. (2016); Archer et al. (2015); Piotrowsky (2012: 74ff); Pilz et al. (2008).

To bypass this difficulty, we initially considered normalizing the Limburgish corpus to Dutch. *Prima facie*, this would seem like a good solution. Dutch is a standardized language and it is known to all Limburgish speakers in the Netherlands and Belgium. It would be relatively easy to find Limburgish staff able to supervise the semi-automatic normalization process from any Limburgish variety into Dutch. Despite these undeniable advantages, we discarded this solution, as introducing Dutch in a Limburgish corpus would have two major drawbacks. First, it would effectively amount to translating the corpus into another language and possibly obfuscate features peculiar to Limburgish. Second, it would rely on an assumption of extreme lexical similarity between Dutch and Limburgish, which a study of the corpus may or may not confirm.

To avoid embedding such assumptions in the design of our corpus, we opted for an alternative strategy. We decided to pick one of the Limburgish varieties as a target for normalization. This was done with the understanding that this would not affect the way other varieties will be represented in the dictionary, but would only facilitate information retrieval in the corpus, mostly for the use of researchers and lexicographers working on the dictionary. Since the largest single-dialect database available to us is the dictionary of the Limburgish Academy Foundation[9], which is easily rendered into contemporary Maastricht-Limburgish, we decided to normalize to the contemporary spelling of the Maastricht dialect. In those cases, where no corresponding Maastricht form exists, a pseudo-Maastricht form will be created on the basis of regular inter-dialectal phonological transformation[10]. To distinguish it from the Maastricht forms attested in the corpus, such pseudo-Maastricht renderings of other dialects will be preceded by an asterisk (*) (see below Table 2).

Elsloo	Roermond	Sittard	Thorn	Valkenburg	Venlo	Weert	Maastricht
spóéze	sjpoeze	sjpoeze	spoeze	sjpoeze	spoeze	spoeze	*spoeze

Table 2: Example of normalization to a pseudo-Maastricht form.

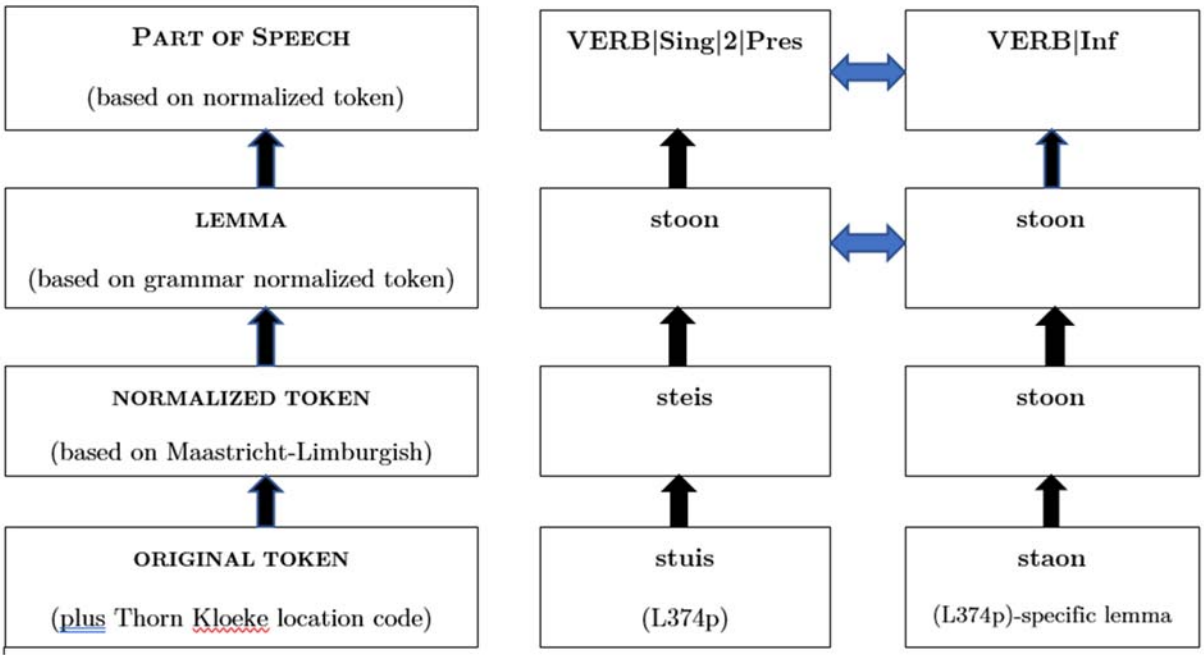
These normalized Maastricht forms will then be added alongside the original dialectal forms, including cases in which the dialectal form is in an idiosyncratic or historical spelling. In the case of an idiosyncratic or historical Maastricht spelling, the form will be paired with a normalized form based on contemporary Maastricht spelling.

[9] See limburgs.org.

[10] Cf. the creation of pseudo-modern forms for historical forms that do not exist anymore in modern languages (e.g. for historical Dutch see Brugman et al., 2016; van Halteren & Rem, 2013).

4.3 Lemmatization and dialect-specific lemma forms

Following our normalization strategy, we will lemmatize the corpus to Maastricht-Limburgish and then tag it for part of speech (PoS-tag) on the basis of grammatical information derived from a Maastricht-Limburgish dictionary. It is important to note that the original tokens will be retained alongside the normalized forms, so that the PoS-tags will be associated with both the Maastricht and the original form (see Table 3). This will allow lexicographers and researchers to analyze the different spellings associated with each lemma and derive dialect-specific lemma forms (see above Table 1 for an example of dialect-specific lemma forms). These dialect-specific forms will eventually feature as headwords in the LCD and enable users to search for and retrieve their preferred spelling of any Limburgish word included in the dictionary (see above Table 1). They will also serve as an indicator of the frequency and distribution of different spelling of a word across Limburg.



PART OF SPEECH (based on normalized token)	VERB|Sing|2|Pres	VERB|Inf
LEMMA (based on grammar normalized token)	stoon	stoon
NORMALIZED TOKEN (based on Maastricht-Limburgish)	steis	stoon
ORIGINAL TOKEN (plus Thorn Kloeke location code)	stuis (L374p)	staon (L374p)-specific lemma

Table 3: General form (left column) and example of normalization, lemmatization, and PoS tagging of a conjugated form (middle) found in a specific dialect and the connection pathway to its dialect-specific lemma (right).

Given the importance of dialect-specific lemma-forms in this project, we initially intended to perform a double lemmatization and pair each token with both its dialect-specific lemma and the corresponding lemma in Maastricht-Limburgish. After much consideration we discarded this approach. In the rest of this section we outline the options we had initially favored and the rationale for choosing a different strategy. We hope that our experience may benefit other projects dealing with the lexicographic representation of regional spelling variation in a corpus.

Initially, we considered relying on existent lexicography and location metadata to pair each token with the corresponding lemma form recorded in dictionaries of the relevant token. This approach presupposes that all words associated with a certain location are amenable to the same lemmatized form, thus not allowing for variant spellings within the dialect. We discarded this idea in favor of a corpus-driven approach which would allow us to derive lemma-forms directly from the corpus and thus account for intra-dialect variation. To this end, we initially aimed to pair each token with a corpus-derived lemma-form that would match the regional spelling of the token. We soon realized that this model, too, was not viable, because it assumes a morphological correspondence between a token and its lemma form. Unfortunately, several Limburgish verbs violate this assumption. For example, in the dialect of Valkenburg the indicative second-person singular of the verb 'to stand' is <sjteis>. At the current state of research, there is no morphological transformation rule that determines whether this form should be matched to <sjtaon> or <sjtoon>, both of which are possible spellings of the infinitive of this verb in this location (see Tables 2 and 3 above). It is possible that predictable transformation patterns for these verbs will emerge from a study of our corpus and make automated lemmatization to a dialect-specific lemma form possible. In the meantime, we will have to dispense with dialect-specific lemmatization and derive lexicographic information on dialect-specific lemma forms only from tokens morphologically identical to the lemma[11]. Thus, the frequency of the Valkenburg lemma form <sjtaon> as opposed to the Maastricht form <stoon> will be calculated on the basis of tokens spelled <sjtaon> only (i.e. the infinitive and indicative first and third person plural), and will not be derived from other conjugated forms. It remains to be determined whether the tokens morphologically identical to the lemma form will constitute a sufficient and reliable indicator of the overall frequency and distribution of a spelling variant. Information about the frequency and distribution of the spelling of other selected conjugated forms (e.g. indicative second person singular or sample past tense forms) may be added to provide a more complete representation of spelling variation across Limburg.

[11] The full verbal paradigm for this verb in Valkenburg-Limburgish based on the local dictionary is the following.

Present tense:		Past tense:	
1s <sjtaon>	1p <sjtoon> / <sjtaon>	1s <sjtóng> / <sjting>	1p <sjtónge> / <sjtinge>
2s <sjteis>	2p <sjtaot>	2s <sjtóngs>	2p <sjtóngt>
3s <sjteit>	3p <sjtoon> / <sjtaon>	3s <sjtóng>	3p <sjtónge> / <sjtinge>

Pp <gesjtange>. Imperative s <sjtank>, p <sjtaot>.

4.4 Considering NLP tools for Limburgish spelling variation

Several software options have been identified for a tentative pipeline. VARD[12] is being considered as a spelling normalizer, Frog[13] for tokenization, lemmatization and PoS-tagging, and Sketch Engine[14] for corpus analysis. Dictionary writing software, such as TshwaneLex[15] and DPS from IDM[16], are also being considered, but will not be further discussed in this article.

At the time of writing this article, testing is still in a very preliminary stage. Only some general comments about the usefulness of VARD and Frog to our project can be made, whereby the focus will be on Limburgish spelling variation.

4.4.1 VARD

VARD was initially built to deal with spelling variation in Early Modern English (Baron & Rayson, 2009), but can potentially be re-trained for other languages[17]. VARD normalizes spelling by inserting a normalized lemma in the place of the spelling variant and retains the original form in an XML tag. VARD can be used in two ways: to manually standardize texts or to automatically standardize a set of texts or corpora (Baron & Rayson, 2009). VARD is a well-known tool and we will not elaborate on it further, except insofar as evaluating it as a potential option for our project.

Since we are still in the process of collecting our corpus, and Limburgish writing exhibits such a high degree of spelling variation, we do not yet know all the variants we will encounter. To gain some preliminary understanding of how much spelling variation we can expect to encounter in our project, we used VARD 2.5.4 for an initial assessment of variation. We first tested diachronic texts from the Maastricht dialect and subsequently synchronic texts in different spellings from the main Limburgish dialect areas for token recognition based solely on a curated word list *before training* VARD. Employing contemporary Maastricht-Limburgish spelling, we created a curated word list for VARD. It contains all parts of speech with inflected forms and consists of 85,731 unique words out of a total of 126,755 words, whereby duplicates existed for separate entries for polysemous words, verbal inflections of the past tense, homonyms and tonal opposites.#

[12] ucrel.lancs.ac.uk/vard/about/.

[13] languagemachines.github.io/frog/.

[14] sketchengine.co.uk.

[15] tshwanedje.com/tshwanelex/.

[16] idm.fr.

[17] For example for historical Dutch (Tjong Kim Sang, 2015), historical Portuguese (Reynaert et al., 2012), and historical German (Pilz et al., 2008).

We tested nine diachronic Maastricht-Limburgish text samples, including literary and Wikipedia texts, in their original spellings spanning the period of ca. 1775–2017. All texts were about 4500 tokens each, except three of the older texts which only have about half as many tokens each. As expected, the percentage of tokens recognized is well over 90% for texts written after 2010. The Wikipedia text samples, although from 2017, registered a recognition percentage of 78.5%. The lower token recognition is at least in part due to more idiosyncratic spellings, unknown proper nouns, foreign script, foreign tokens, more specialized compounds, and typos. For 20th-century texts, token recognition was 75–85%. Surprisingly, for 19th-century and older texts 45–60% of tokens are still recognized.

For the second test on the same Maastricht-Limburgish texts, replacement rules were added to VARD for spelling phenomena that affected most texts. Baron and Rayson (2009) indicate that VARD's user-defined list of letter replacement rules to compute alternative forms results in a significant increase in performance when automatically normalizing the corpus. These replacement rules for Maastricht-Limburgish included replacements for spelling changes made in 2004 and some 19th-century spelling peculiarities. Some of these rules will also benefit token recognition for many East Limburgish spellings, as these were closer to the Maastricht spelling before the 2004 spelling change. The results of the second test enhanced token recognition on average by about four percentage points, whereby texts from the 21st century gained 2.1%, 20th-century texts 5.7% and pre-1900 texts 4.6%.

Subsequently, we tested nine synchronic text samples from Wikipedia of about 2000 tokens each from all main Limburgish dialect areas[18]. We first used the same approach as mentioned above for the first VARD test. Token recognition for non-Maastricht spellings had a mean of 45%. The range was between 37% for the spelling of the Kerkrade Ripuarian dialect and 56% for the spelling of the Valkenburg East Limburgish dialect, which is geographically close to Maastricht. The results for the second test, with the replacement rules indicated above, resulted in a mean recognition of 51%. There was a range of about 40% for the spelling of the Kerkrade dialect to 62% for the spelling of the Valkenburg dialect.

The results from the diachronic Maastricht texts and the synchronic texts from all main dialect areas can be interpreted as indicators of the different levels of spelling

[18] These included the following dialects: Alken* (West Limburgish), Geleen (East Limburgish), Heerlen (East Limburgish Ripuarian transition area), Kerkrade* (Ripuarian), Montfort (East Limburgish), Ool (East Limburgish), Roermond (East Limburgish), Valkenburg (East Limburgish), Venlo (Mich Quarter transition area). Those with an asterisk (*) only had about half of the tokens.

variation in Limburgish. Considering the fact that this is a pre-trained version of VARD, these results are encouraging. We are contemplating to test and train VARD on a large corpus, which, according to Baron and Rayson (2009: 9), should allow it to better find and rank candidate equivalents for variants found in the remainder of the corpus.

We are considering the following steps regarding VARD. We shall start by training VARD and creating a Maastricht-Limburgish word list that is as extensive as possible. This is crucial, since we normalize to the contemporary spelling of this dialect. We will start with contemporary texts in the Maastricht spelling and subsequently process all Maastricht texts diachronically. Thereafter we intend to process texts in spellings from other dialects. On the basis of a mapping of Limburgish spelling variation we are examining whether to first process texts from dialects with spellings closest to Maastricht-Limburgish, followed by texts that in terms of spelling are progressively farther removed. For each dialect we will first normalize contemporary texts followed by increasingly older texts. Finally, on the basis of a mapping of Limburgish spelling variation, we will also determine whether to create a more extensive list of replacement rules. Some replacement rules to normalize to the Maastricht spelling are common to all dialectal spellings. For the spelling of some (groups of) dialects we might have to create a separate set of replacement rules. Depending on how extensive these separate replacement rules are for different (groups of) dialects we are contemplating training separate VARD applications.

One last issue we need to resolve is how to disambiguate homographs with different meanings in different dialects. Since Limburgish spelling is phonological and the normative spelling tags a grapheme with a particular phoneme, in some instances a word spelled according to the phonology of one dialect exists in another dialect, but with a different meaning. For example, <eur>, a possessive pronoun in Maastricht dialect meaning 'your' (singular polite form and plural), is the possessive pronoun for 'her' in the Venlo dialect. The Maastricht form for 'her', to which it has to be normalized, is <häör>. In a Venlo text, the Maastricht-trained VARD will recognize the token, but will not recognize that it is a variant spelling. Further experimentation will be required to optimize for the tool's maximum effectiveness in normalizing the spelling variation in Limburgish texts. That optimization might include forking the normalization rules for individual dialects, or dialect areas, to force certain normalizations that are specific to only that dialect.

4.4.2 Frog

Frog is a Natural Language Processing suite originally developed for standard Dutch (Van den Bosch et al., 2007). It integrates a series of modules including a tokenizer, lemmatizer, morphological segmenter, and a PoS tagger. It also includes a named entity recognizer, phrase chunker, and dependency parser, but it is still to be determined whether these tools are useful for our project. Frog is originally intended to work on modern standard Dutch, but has been used amongst others by the Nederlab project for historical Dutch (Brugman et al., 2016: 1279). It also contains Froggen, a trainer module part of Toad[19], that allows Frog to be trained for another language. However, Frog relies on a standard spelling to perform its analysis and is not equipped to deal with rich spelling variation. Normalizing Limburgish spelling variation by a pre-processing tool like VARD is therefore a prerequisite.

When evaluating Frog's usefulness for our pipeline, we first need to consider the output content and format from VARD. VARD can create two output formats. One is a version of the text with fully normalized spelling. Another version is the normalized text with XML tags, each encapsulating the original token along with the details of the normalization. As Frog cannot parse the VARD XML output out-of-the-box, we have a few pathways to experiment with to determine which is most compatible and without data loss.

Since Frog does not natively deal with spelling variation we have had to investigate options how to preserve the data of both the text in original spelling and the normalized version. One option is to configure Frog so that it can accept the pseudo-XML that VARD produces. This seems feasible as one of Frog's native formats is an XML format, namely FoLiA XML[20]. We are investigating whether it is possible to adapt Frog's parser to read VARD's pseudo-XML format. This, potentially, would allow us to preserve the connection between original token and normalized token through Frog's processing. Another, possibly simpler, option would be to insert an original token column to Frog's tab-delimited output of processed normalized tokens. This means that only normalized tokens are present in Frog's processing, but the connection to the original text is re-established in a secondarily, post-Frog processed output file.

We will not consider here in depth the steps in the pipeline after this point. However, one possible Frog output option is a tab-delimited text file, which Sketch Engine can process. The content of the Frog output will certainly include token, lemma, and PoS columns. Additional output from Frog may be included, depending on Sketch Engine's ability to parse the information and include it in its word sketches. Finally, header information, including items like location code, date, and author information, will be appended to the file to be read into Sketch Engine. At this point we will have attempted to preserve all the data from the source texts including all the tagging, and the corpus would be ready for analysis in Sketch Engine.

[19] github.com/LanguageMachines/toad/releases/tag/v0.3.

[20] For FoLiA XML see van Gompel and Reynaert (2013).

5. Conclusion

In this paper, we introduced a new project at Maastricht University for the creation of a Limburgish Corpus Dictionary (LCD). Limburgish spelling variation, diachronic spelling data, writing practices and language policy present us with the possibility to look for novel ways to process and display this non-standardized regional language. We first presented a model of how to display the spelling variation in Limburgish in an online dictionary, based on how Limburgers use their language and the policy to treat all dialects and spelling variation equally. For NLP processing purposes we then discussed the reasons to use the Maastricht-Limburgish variety as a normalizing standard. We also developed a set of heuristics to retrieve dialect-specific forms that will eventually feature as headwords in the LCD. This will enable users to search for and retrieve 'their' preferred spelling of any Limburgish headword included in the dictionary from the myriad of spellings that a Limburgish lemma can have. The dialect-specific forms will also serve as an indicator of the frequency and distribution of different spellings of a lemma across Limburg. We then discussed possible software options for a tentative pipeline and the steps we consider taking to further investigate their usefulness for our project. Our focus will now be on determining how the available NLP software options will allow us to execute our project in conformity with the lexicographic model we have developed for Limburgish. This will enable us to present Limburgish-speakers with a free online dictionary that represents their real language usage.

6. References

Archer, D., Kytö, M., Baron, A. & Rayson, P. (2015). Guidelines for normalising Early Modern English corpora: Decisions and justifications, *ICAME Journal*, 39, pp. 5-25.

Bakkes, P., Crompvoets, H., Notten, J. & Walraven, F. (2003). *Spelling 2003 voor de Limburgse dialecten.* Available at: http://www.limburgsedialecten.nl/download/spelling2003.pdf.

Baron, A. & Rayson, P. (2009). Automatic standardization of texts containing spelling variation. How much training data do you need? In M. Mahlberg, V. González-Díaz and C. Smith (eds.) *Proceedings of the Corpus Linguistics Conference*, CL2009, University of Liverpool, UK, 20-23 July 2009.

Barteld, F., Schröder, I. & Zinsmeister, H. (2016). Dealing with word-internal modification and spelling variation in data-driven lemmatization, *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pp. 52–62.

Belemans, R. (2009). *Van (Limburgse) dialecten naar Europees erkende streektaal en/of immaterieel cultureel erfgoed? De invloed van nationale taalpolitiek en van internationaal erfgoedbeleid op de perceptie van en op de overheidszorg voor endogene taalvariatie in Vlaanderen.* Doctoral dissertation Universiteit Leuven.

Belemans, R., Kruijsen, J. & van Keymeulen, J. (1998). Gebiedsindeling van de zuidelijk-Nederlandse dialecten. *Taal en Tongval*, 50, pp. 25-42.

Boersma, P. (2013). *The history of the Franconian tone contrast.* Available at: http://www.fon.hum.uva.nl/paul/papers/FranconianToneHistory68.pdf.

Brugman, H., Reynaert, M., van der Sijs, N., van Stipriaan, R., Tjong Kim Sang, E. & Van den Bosch, A. (2016). *Proceedings of LREC*, Portoroz: ELRA, pp. 1277-1281.

Daan, J. & Blok, D. (1969). Van Randstad tot Landrand. In *Bijdragen en Mededelingen der Dialectcommissie van de KNAW XXXVI.* Amsterdam: Noord-Hollandsche Uitgevers Maatschappij.

Draye, L. (2007). Enkele klank- en vormkenmerken van de Limburgse dialecten. In R. Keulen, T. van de Wijngaard, H. Crompvoets & F. Walraven (eds.). *Riek van Klank; Inleiding in de Limburgse dialecten.* Sittard: Veldeke Limburg, pp. 24-44.

Fournier, R., Verhoeven, J., Swerts, M. & Gussenhoven, C. (2004). Prosodic and segmental cues to the perception of grammatical number in two Limburgian dialects of Dutch. *Proceedings of the Speech Prosody 2004 Conference*, pp. 713-716.

Gussenhoven, C. (2007). De Limburgse tonen. In L. Heijenrath & S. Kroon (eds.), *Jaarboek 2006.* Roermond: Veldeke Limburg, pp. 21-32.

Gussenhoven, C. & Peters, J. (2008). De tonen van het Limburgs. In *Nederlandse Taalkunde*, 13, pp. 87-114.

Hendrickx I. & Marquilhas, R. (2011). From Old Texts to Modern Spellings: An Experiment in Automatic Normalisation, *Journal for Language Technology and Computational Linguistics*, 26(2), pp. 65-76.

Hoppenbrouwers C. & Hoppenbrouwers G. (2001). *De indeling van de Nederlandse streektalen, Dialecten van 156 steden en dorpen geklasseerd volgens de FFM.* Assen: van Gorcum.

Jongen, L. (2016). Van het begin tot 1500. Van geschreven naar gedrukte letters. In L. Spronck, B. van Melick & W. Kusters (eds.) (2016). *Geschiedenis van de literatuur in Limburg.* Nijmegen: Uitgeverij Vantilt, pp. 23-65.

Kestemont, M., Daelemans, W. & De Pauw, G. (2010). Weigh your words - memory-based lemmatization for Middle Dutch, *Literary and Linguistic Computing*, 25(3), pp. 287-301.

Leerssen, J.Th., Crompvoets, H., Walraven, F., Segers, J., Belemans, R., Bakkes, P. & Gillessen, L. (1996). *Verslag werkgroep erkenning Limburgs als streektaal.* Available at: http://jonckbloet.hum.uva.nl/leerssen/images/limburgs/adindex.html.

Opgenort, J. R. (2012). *Limburgse taal.* Available at: http://www.opgenort.nl/limburgse_taal.

Pilz, Th., Ernst-Gerlach, A., Kempken, S., Rayson, P. & Archer, D. (2008). The identification of spelling variants in English and German historical texts: manual or automatic?, *Literary and Linguistic Computing*, 23(1), pp. 65-72.

Priotrowski, M. (2012). *Natural Language Processing for Historical Texts.* San

Rafael: Morgan & Claypool.

Reynaert, M. (2014). TICCLops: Text-Induced Corpus Clean-up as online processing system, *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, Dublin, Ireland, August 23-29, pp. 52-56.

Reynaert, M., Hendrickx, I. & Marquilhas, R. (2012). Historical spelling normalization. A comparison of two statistical methods: TICCL and VARD2, *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2)*, Lisbon, pp. 87-98.

Robinson, O.W. (1992). *Old English and its closest relatives; A survey of the earliest Germanic languages.* Stanford: University Press.

Schutter, de, G. & Hermans, B. (2013). The Limburg dialects: Grammatical properties. In F. Hinskens & J. Taeldeman (eds.) *Language and Space. An International Handbook of Linguistic Variation*, 3. Berlin/Boston: De Gruyter Mouton, pp. 356-377.

Souvay, G. & Pierrel, J.-M. (2009). LGeRM Lemmatisation des mots en Moyen Français. Traitement Automatique des Langues, *ATALA*, 50(2), pp. 149-172.

Spronck, L. (2018). *Van Sermoen tot Percessie. Het Maastrichts rond 1800* (to be published).

Spronck, L. (2016). 1793-1893. Verandering van het blikveld: verlies en winst. In L. Spronck, B. van Melick, & W. Kusters (eds.) (2016). *Geschiedenis van de literatuur in Limburg.* Nijmegen: Uitgeverij Vantilt, pp. 203-309.

Spronck, L. (1962). De Maastrichtse dialektliteratuur voor 1840. In Miscellanea Trajectensia; *Bijdragen tot de geschiedenis van Maastricht*, Werken LGOG nr. 4, Maastricht: LGOG, pp. 435-495.

Spronck, L., Salemans, B. & Schrijnemakers, S. (2007). Maastricht: Het Maastrichts anno 1807: boers? de gelijkenis in het Maastrichts besproken. In F. Bakker & J. Kruijsen (eds.), *Het Limburgs onder Napoleon. Achttien Limburgse en Rijnlandse dialectvertalingen van 'De verloren zoon' uit 1806-1807.* Utrecht: Gopher, pp. 177-215.

Swanenberg, J. (2013). All dialects are equal, but some dialects are more equal than others, In *Tilburg Papers in Culture Studies*, Paper 43.

Tervooren, H. (2005). *Van der Masen tot op den Rijn; ein Handbuch zur Geschichte der mittelalterlichen volkssprachlichen Literatur im Raum von Rhein und Maas.* Band 105. Geldern: Historisches Verein für Geldern und Umgegend.

Tjong Kim Sang, E. (2015). *Converting seventeenth century Dutch to modern Dutch.* Presented at the Workshop Morphosyntactic Enrichment of Historical Texts, Utrecht, The Netherlands.

Van den Bosch, A., Busser, G., Canisius, S. & Daelemans, W. (2007). An efficient memory-based morpho-syntactic tagger and parser for Dutch. In P. Dirix et al. (eds.), *Computational Linguistics in the Netherlands: Selected Papers from the Seventeenth CLIN Meeting*, Leuven, Belgium, pp. 99-114.

van Gompel, M. & Reynaert, M. (2013). FoLiA: A practical XML format for linguistic annotation: A descriptive and comparative study. *Computational Linguistics in the Netherlands Journal*, 3, pp. 63-81.

van Halteren, H. & Rem, M. (2013). Dealing with orthographic variation in a tagger-lemmatizer for fourteenth century Dutch charters, *Language Resources and Evaluation*, 47, pp. 1233-1259.

van Horen, H. & van Horen-Hoosel, H. (2016). 1500-1793. Duistere eeuwen? In L. Spronck, B. van Melick & W. Kusters (eds.), *Geschiedenis van de literatuur in Limburg*. Nijmegen: Uitgeverij Vantilt, pp. 66-199.

van Hout, R. (2007). Het Europese Handvest en het Limburgs: het politieke en taalkundige discours. In H. Bloemhoff & P. Hemminga (eds.), *Streektaal en duurzaamheid. Lezingen van de internationale streektaalconferentie in Noordwolde*, 25 mei 2007, Berkoop/Oldeberkoop: Stichting Stellingwarver Schrieversronte, pp. 33-47.

van Hout, R., & Münstermann, H. (1981). Linguistische afstand, dialect en attitude. *Gramma: Nijmeegs Tijdschrift voor Taalkunde*, 5(2), pp. 101-123.

Weijnen, A., Goossens, J. & Goossens, P. (1983). *Woordenboek van de Limburgse dialecten. Inleiding & I. Agrarische terminologie*. Aflevering 1. Assen: Van Gorcum.

Weusten, S., Grondelaers S. & Van Hout, R. (2013). De herkenning en waardering van zes Limburgse 'stadse' dialecten. Leve het Maastrichts? In P. Bakkes (ed.), *Jaarboek Veldeke Limburg*, 20. Roermond: Vereniging Veldeke Limburg, pp. 61-75.

Willemyns, R. (2003). *Het verhaal van het Vlaams; De geschiedenis van het Nederlands in de Zuidelijke Nederlanden*. Antwerpen: Standaard Uitgeverij.

Wolters, L. (2016). *Veldeke Limburg 1926-2016*. Roermond: Veldeke Limburg.

Websites:

e-wld.nl. Accessed at: e-wld.nl. (20 April 2017)

github.com/LanguageMachines/toad/releases/tag/v0.3. Accessed at: https://github.com/LanguageMachines/toad/releases/tag/v0.3. (20 March 2017)

idm.fr. Accessed at: http://www.idm.fr/. (27 March 2017)

languagemachines.github.io/frog/. Accessed at: https://languagemachines.github.io/frog/ (27 March 2017)

limburghuis.nl. Accessed at: https://limburghuis.nl/. (20 April 2017)

limburgs.org. Accessed at: http://www.limburgs.org/en/limburgish. (26 March 2017)

limburgsewoordenboeken.nl. Accessed at: www.limburgsewoordenboeken.nl. (20 April 2017)

meertens.knaw.nl/kloeke. Accessed at: https://www.meertens.knaw.nl/kloeke/. (22 April 2017)

mestreechtertaol.nl. Accessed at: http://www.mestreechtertaol.nl/dictionair/mst. (20 April 2017) *sketchengine.co.uk*. Accessed at: https://www.sketchengine.co.uk/. (27 March 2017)

tshwanedje.com/tshwanelex. Accessed at: http://tshwanedje.com/tshwanelex/. (27 March 2017)

ucrel.lancs.ac.uk/vard/about. Accessed at: http://ucrel.lancs.ac.uk/vard/about/. (22 March 2017).

woordenboek.gronsveld.com. Accessed at: woordenboek.gronsveld.com. (20 April 2017)