Identification of Languages in Linked Data: A Diachronic-Diatopic Case Study of French

Sabine Tittel[1], Frances Gillis-Webber[2]

[1] Heidelberg Academy of Sciences and Humanities, Seminarstraße 3,
D–69117 Heidelberg, Germany
[2] Department of Computer Science, University of Cape Town, Cape Town, South Africa
E-mail: sabine.tittel@urz.uni-heidelberg.de, fran@fynbosch.com

Abstract

When modelling linguistic resources as Linked Data, the identification of languages using language tags and language codes is a mandatory task. IETF's BCP 47 defines the standard for tags, and ISO 639 provides the codes. However, these codes are insufficient for the identification of diatopic variation within a language and, also, for different historical language stages. This weakness hampers the accurate identification of data, which in turn leads to ambiguity when extending, aggregating and re-using this data—a key notion of Linked Open Data and the Semantic Web. We show the limitations of language identification with a case study of French linguistic data from both a diachronic and a diatopic perspective. Our exemplary data derives from dictionaries of Old French, Middle French, and of Modern French dialects, and from a Modern French linguistic atlas. For each exemplar, we propose a solution using the *privateuse* sub-tag of BCP 47's language tag, staying within the boundaries of existing standards. Using a predefined pattern for the *privateuse* sub-tag, the solutions enable a dialect, a patois, in combination with a time period, to be defined and identified. This can lead to shared agreement of language tags that will increase interoperability within the context of Linked Data.

Keywords: language codes; language tags; language annotation; Linked Open Data; French dialects

1. Introduction

Over the last decade, modelling linguistic data using the Resource Description Framework (RDF), following the Linked Data (LD) paradigm, has become a widespread method for the creation of datasets for a multilingual web of data. It enables machine-readable, cross-resource access to data that are otherwise spread across the web as isolated datasets. However, for the modelling of linguistic resources as LD, the use of language tags is essential: the annotation with language tags whose form adheres to established standards ensures unambiguous language identification of linguistic information, such as lexemes and their graphic and phonetic realizations. Because of the interlinking of lexemes and their different realizations, the LD format can be particularly valuable for linguistic resources that document the diatopic diversity of a given language (i.e., with a spatial reference). Examples are regional dictionaries or linguistic atlases. These resources can be complemented with historical data to introduce a diachronic perspective to the diatopic variation of the language (i.e., considering evolution through history). This can be, e.g., data from historical dictionaries that indicate regional characteristics. The publication of these resources as LD and the corresponding means of data query can enhance studies that focus on the diatopic richness of modern-day languages and on the evolution of diatopic variation at the same time. The use of language tags is specified by IETF's BCP 47 (Phillips & Davis, 2009: 1-4) and the required language codes come from ISO 639 (International Organization for Standardization, n.d.). Within our field, however, we observe a lack of language tags and codes hampering the required language annotation. In this paper, we address the issue of language tagging with French linguistic resources combining a diatopic with a diachronic perspective: in a case study, we investigate data of Old-, Middle- and Modern French resources with (regional) dictionary data and linguistic atlas data.

After a short outline of the diachronic-diatopic landscape of French linguistic resources (Section 1.1), we briefly describe RDF, LD (Section 1.2), and the identification of languages (Section 1.3). In the following section, we introduce the use of a pattern for language tags (Section 2). Our case study of French uses exemplary data of historical and modern dictionaries (Section 3) and of a linguistic atlas (Section 4). For each exemplar, we demonstrate a solution for the language tagging, using the pattern described. We evaluate the solutions in Section 5, and in Section 6, we present an interface which can be used to generate (and decode) language tags according to our pattern. We conclude the paper in Section 7.

1.1 Diatopic linguistic resources and a diachronic perspective

The regional varieties, dialects and patois[1] of the French of France are under-represented in linguistic consideration in general and in lexicography in particular (Rézeau, 2001: 7). This is all the more true for the diatopic reflection from a diachronic angle: the historical development of French regionalisms has not been studied in a comprehensive yet detailed way (Gleßgen & Thibaut, 2005: XII). Studies focusing on single topics such as a particular region in a particular time period have been conducted, recently by, e.g., Chauveau (2016), and Rézeau (2016).

There are many resources that can be exploited for diatopic-diachronic studies: for the different language periods of French, dictionaries, corpora, and—for modern French in

[1] We are aware of the discussion of the terms that denote different variations within the diatopic diasystem of French. In this paper, we will use the terms following the French literature, where *régionalité linguistique* (of French) is clearly distinguished from *dialectes*, the first referring to variation within the standard language, the latter to the primary dialects of France that are the successors of the Old French dialects (Gleßgen & Thibaut, 2005: V), and patois typically designating a local variety of a dialect. Note that we use 'patois' as a non-pejorative term.

particular—linguistic atlases are available.[2] Modern resources covering French varieties include dialect or patois dictionaries (e.g., Rézeau, 2001; Varlet, 1896; Vasseur, 1998), linguistic atlases (e.g., Gilliéron & Edmont, 1902–1910; Lanher et al., 1979–1988; Dondaine & Dondaine, 1972–1991), corpora (Thun, 2011)[3], and, also, individual studies (e.g., Rézeau, 2007) focusing on regional French, dialects and patois. For the historical language stages however, there are fewer resources with diatopic content. A reason for this is that from ca. 1500 AD—with the constitution of French (evolving from a Parisian scripta[4] that had occurred around 1250) as a national language (Wolf, 1979: 94f.)—to the beginning of the 19th century, dialects almost exclusively belonged to the oral culture (Berschin et al., 2008: 203–211). Consequently, studies on the subject of regionalisms are scarce for this time period. Earlier however, in medieval times, the primary dialects included in the notion of Old- and Middle French, such as Picard and Anglo-Norman, were used for both oral and written communication. Hence, we look at the transmission of numerous linguistic primary resources (texts in manuscripts, often accessible in scholarly text editions) documenting regional variation during the Middle Ages. For this time period, studies mainly focus on a single primary resource and how to localize its language in a specific region (notably works by J.-P. Chambon, e.g., Chambon, 1997, and G. Roques, *cf.* the 'Liste Roques' in Glessgen & Trotter, 2016: 473–635). There are also many-volumed, comprehensive dictionaries of the historical language stages, in particular the *Dictionnaire étymologique de l'ancien français* (DEAF, Baldinger et al., 1971–) for Old French, the *Dictionnaire du moyen français* (DMF, ATILF – CNRS & Université de Lorraine (2015)) for Middle French, and the *Französisches Etymologisches Wörterbuch* (FEW, von Wartburg, 1922–) for the diachronic description of French until the present day. These dictionaries—although not necessarily conceived as data sources for diatopic linguistics—provide a synopsis of the knowledge of the particular historical language stage. By incorporating the results of historical dialect studies, they thus contribute to our knowledge of regional variation evolving through time.

Digitization of diatopic resources. It is a European consensus that geographic variation of languages needs to be valorized and promoted, particularly online: UNESCO, La Francophonie[5] and other international organizations emphasize the need for (culturally and) linguistically diverse local content to be published online and for a vitalization of multilingualism on the Web, *cf.* Vannini & Le Crosnier, 2012: 13–21. A large number of the resources in our focus—word lists, dictionaries, linguistic atlases, texts—are currently only available in print. Only a few are available in digital form, and mostly

[2] We identified five language periods of French, *cf.* Gillis-Webber et al. (2019: Section 4 with Fig. 4).

[3] Corpus of letters written by prisoners, soldiers, prostitutes, etc., that document the diatopic variation within the French substandard language.

[4] The written form of a spoken dialect.

[5] https://www.unesco.com/; https://www.francophonie.org/ [13-02-2019].

as digital images.[6] Many have yet to be (retro-)digitized. Digitization would allow for "many new approaches to the quantitative comparison of languages, be it for a better understanding of cross-linguistic variation in grammatical structure or for new and improved historical comparative reconstructions" (Bouda & Cysouw, 2012: 15). One such approach is the representation of the resource in RDF, which in turn allows for the extension to LD.

1.2 Enabling resource integration with the Resource Description Framework and Linked Data

RDF[7] is a data model that represents knowledge in a graph data structure facilitating data interchange on the (Semantic) Web. It is a fundamental technology of the Semantic Web, in which data is structured and meaning can thus be inferred by machines. RDF expresses data as sets of statements in the form of *subject-predicate-object*-triples. Each *subject* and *object* is a node; the *predicate* (or *property*) forms a relation (edge) pointing from the source node (*subject*) to a target node (*object*). Nodes and edges are identified with URIs (Uniform Resource Identifier, accessible via HTTP), and the object can also be described as a string literal (Cyganiak et al., 2014). LD can be described as a set of recommended practices for publishing RDF as structured data on the Web (Bizer et al., 2009). Applying LD principles (Berners-Lee, 2006) to the modelling of linguistic data comes with significant advantages, such as structural interoperability (cross-resource access by using same format and same query language), conceptual interoperability (through shared vocabularies), accessibility (through standard Web protocols), and resource integration by means of interlinking (Chiarcos et al., 2013). Because of the exploratory nature of LD, URIs identifying, e.g., lexemes, their senses, and their concepts referring to the things denoted, *things* and the usage of their *designations* can be explored in a cultural context without being restricted to the vehicle of a particular language. The integration of resources of different language stages and diatopic variation enables observation through time and space, including, e.g., borrowing and word formation processes, and semantic shift within a large data collection. For Old French, the first steps have been made by modelling exemplary lexicographic data of the DEAF as LD using the OntoLex-Lemon vocabulary[8], and the modelling of a scholarly text edition of a Middle French medical treatise using RDFa (Tittel & Chiarcos, 2018; Tittel et al., 2018). To the best of our knowledge, there are no other historical linguistic resources of French represented as LD that could be exploited for diachronic-diatopic studies.

[6] *Cf.*, e.g., the references at https://www.lexilogos.com/lorrain_dictionnaire.htm [10-06-2019].

[7] RDF 1.1. Primer, 2014, https://www.w3.org/TR/rdf11-primer/ [10-05-2019].

[8] https://www.w3.org/2016/05/ontolex/ [13-05-2019].

1.3 Identification of languages

When modelling linguistic resources in RDF, it is necessary to identify the language of the resource and the information therein (be it a *word*, a *multiword expression*, a *sense*, a *graphical realization*, a *phonetic representation*), and to annotate literals with a language tag. IETF's BCP 47 specifies the Best Current Practice for language tags; the language tag typically begins with a language code and it must conform to established standards (Cyganiak et al., 2014). The language code comes from external resources such as ISO 639, which provides the authoritative list of language codes. Alternatives are catalogues like Glottolog, Ethnologue, and MultiTree.[9] However, these alternatives do not meet the requirements of BCP 47 for the encoding of languages. They also reveal significant shortcomings concerning registration, hierarchization, diachronic and dialectal criteria, all of which have been discussed in detail in Gillis-Webber and Tittel (2019: 4:6-8) and Gillis-Webber et al. (2019). Lexvo[10] provides dereferenceable URIs only for languages registered by ISO 639 (de Melo, 2015). It is, thus, insufficient for our use.

An exemplary lexical entry in RDF (identified as **E0**), modelled using OntoLex-Lemon and serialized in Turtle[11] is:

```
@PREFIX :         <http://www.example.com/entry/> .
@PREFIX ontolex:  <http://www.w3.org/ns/lemon/ontolex#> .
@PREFIX lexinfo:  <http://www.lexinfo.net/ontology/2.0/lexinfo#> .
@PREFIX dct:      <http://purl.org/dc/terms/> .
@PREFIX rdfs:     <http://www.w3.org/2001/02/rdf-schema#> .
@PREFIX dbpedia: <http://www.dbpedia.org/resource/> .

:alconorque a ontolex:LexicalEntry , ontolex:Word ;
    lexinfo:partOfSpeech lexinfo:Noun ;
    dct:language      <http://lexvo.org/id/iso639-1/pt> ,
                      <https://iso639-3.sil.org/code/por> ;
    rdfs:label         "cork oak"@en , "alconorque"@pt ;
    ontolex:denotes   dbpedia:Quercus_suber .
```

[9] https://glottolog.org, https://www.ethnologue.com, http://multitree.org/ [07-06-2019].

[10] http://lexvo.org [07-06-2019].

[11] Terse RDF Triple Language, http://www.w3.org/TR/turtle/ [11-01-2019]. In the following code examples, namespaces are assumed defined the usual way. We include hypothetical URIs using the namespace <http://www.example.com/entry/>.

where Lines 10-11 show the applicable language URIs for the lexical entry indicated as ‘Portuguese’, from ISO 639-1 and ISO 639-3 respectively; Line 12 shows the language code ‘English’ (ISO 639-1 ‘en’) for the literal “cork oak”, and the language code ‘Portuguese’ (ISO 639-1 ‘pt’) for the literal “alconorque”.

The ISO 639 standard shows significant shortcomings with respect to regional variation and to historical language stages, as was shown in Gillis-Webber and Tittel (2019: 4:4-5); *cf.* also Figures. 4 and 5. This prevents the unambiguous identification of languages, even more so when modelling multiple ‘snapshots’ of data of the same language through time and space.

2. Pattern for Language Tags

As demonstrated in **E0**, the use of ISO 639 language codes in language tags is straightforward for most modern and well-known languages. However, the problem of missing or inadequate language codes extends to any variety or dialect of a language which is requires representation on the web, and for which an ISO 639 code is simply not available. Language tags, as prescribed by BCP 47, have the syntax:

language-extlang-script-region-variant-extension-privateuse

with each portion, called a sub-tag, separated by a hyphen (Phillips & Davis, 2009: 4). Gillis-Webber & Tittel (2019) propose a pattern for the *privateuse* sub-tag.[12] The pattern for the *privateuse* sub-tag is of the form:

x-language-otherlect-timeperiod-region-uri

where x- is a BCP 47 requirement indicating *privateuse*, and language (a language, dialect, patois or pidgin), otherlect (an ethnolect, sociolect, or idiolect), timeperiod, region, and URI are all parts of the sub-tag, separated by a hyphen (Gillis-Webber & Tittel, 2019: 4:12). Apart from the *privateuse* sub-tag, the sub-tags are specified by BCP 47 as “identified on the basis of its length, position in the tag, and its content”; each sub-tag typically is part of an ISO standard or registry (*ib.*) For the *privateuse* sub-tag, the use of a key (Table 1) is proposed to identify each part, thus allowing for flexibility of content and variable length thereof.

[12] Note that this pattern is not intended to replace any content that would typically be included in other sub-tags. To see the most recent updates to the pattern, please go to: https://londisizwe.org/ language-tags/.

Part	**Key 1**	**Key 2**
language	0	0 = User-defined 1 = Glottocode
otherlect	1	0 = User-defined 1 = Glottocode
timeperiod	2	0 = one year only, BC 1 = one year only, AD 2 = start:BC - end:BC 3 = start:BC - end:AD 4 = start:AD - end:AD
region	3	0 = Geohashed latitude and longitude coordinates – polygon 1 = Geohashed latitude and longitude coordinates – point only 2 = URI to GeoJSON-LD 3 = Code from ISO 3166 4 = Identifier from GeoNames
URI	4	0 = URI shortcode from https://londisizwe.org/language-tags/

Table 1: The key for each part in the *privateuse* tag.

We identified the following set of competency questions (CQs) for the pattern, where [lect] can be replaced by any language, variant, dialect, patois, and scripta.

CQ 1 How to identify a [lect] that has no ISO 639 language code, but whose parent language does?

CQ 2 How to identify a [lect] for which ISO 639 provides a language code that indicates a different time period?

CQ 3 How to identify a [lect] for which ISO 639 provides two language codes?

CQ 4 How to identify a [lect] in space that has neither an ISO 639 code nor a code from an alternative directory?

CQ 5 How to identify a [lect] in time?

CQ 6 How to identify endonyms and exonyms of a [lect]?

When evaluating the pattern, these CQs should be answerable. Using the case study of French, we will revisit the CQs in Section 5 to test the efficacy of the proposed pattern.

3. Modelling of regional variation in dictionary data

For our case study, we will embrace both diachronic and diatopic data of French, with the latter typically mirroring aspects of the former.

3.1 Old French

Old French should be understood as an umbrella term for a number of dialects resulting from the process of settlement and romanization, different substrates, strates, etc. These dialects present distinctive linguistic realities from the beginning of the 12th century, *cf.* Rickard (1974: 54–65; 71–84).

For the Old French period, the contribution of the DEAF to our knowledge of diatopic variation of Old French has been discussed by Möhren (2016) and Tittel (2016). The DEAF allows for the annotation of data with 35 scriptae, including broader categories like 'Nord-Est' or 'Centre' (*cf.* Figures 4 and 5). For Old French, the ISO 639-3 language code is 'fro' («842–ca.1400»), but there are no ISO 639 language codes available for the scriptae except for Anglo-Norman ('xno') and Judéo-French ('zrp'). For the modelling of DEAF data with OntoLex-Lemon, although 'fro' has been used as the language tag, this does not allow for the data to be differentiated on scriptae (Tittel & Chiarcos, 2018: 64f.).

An exemplar (**E1**) derived from the DEAF is *jannaie* (designating a terrain covered with gorse), a lexeme marked as Gallo.[13] It can be modelled as follows:

```
1 :jannaie a ontolex:LexicalEntry , ontolex:Word;
2 ontolex:canonicalForm :jannaie_lemma .
3
4 :jannaie_lemma a ontolex:Form ;
5 ontolex:writtenRep "jannaie"@fro-x-00gallo .
```

In our language tag on Line 5, as an ISO 639 language code does not exist for (Old) Gallo, we have made use of a compiled language tag: fro identifies it as from the Old French period, and 00 indicates that it is a user-defined language (i.e., a code from an alternative directory to ISO 639 has not been used).[14]

[13] DEAF J 136,9; https://deaf-server.adw.uni-heidelberg.de/lemme/jaon#jannaie [10-05-2019].

[14] For a discussion of further approaches to language tagging Old French dialects, *cf.* Gillis-Webber & Tittel (2019: 4:9-11).

3.2 Middle French

The comprehensive dictionary for the Middle French period is the DMF. With respect to the study of dialectal characteristics of the Middle French lexis, the DMF is a resource of limited value and difficult access (Renders, 2016: 95f.). However, the DMF has the potential for facilitating the study of diatopic variation of late medieval French: the data structure of the DMF entry does not contain a label that specifically tags information as being dialectal (thus, the information cannot easily be accessed in a machine-aided way), but the running (unstructured) text of approx. 1,190 entries (Renders, 2016: 89) includes in effect such information; this can be exploited.

Although the French written standard spread in Middle French time, the dialects still maintained their role in the literature. The DMF defines a list of 29 "étiquettes régionales" (Renders, 2016: 86) comparable with the DEAF scriptae list. For Middle French, the ISO 639-3 language code is 'frm' («ca. 1400–1600»); this can be utilized to identify the language, but the challenge of codes for its dialects needs to be addressed.

In the following exemplar (**E2**), we model a lexeme that is marked as dialectal: *appreper* v. "s'approcher (d'un lieu)" "Région. (Wallonie)".[15] The language code from ISO 693-1 for modern Walloon is 'wa'. But as for the Old French language period, the code should not be used for the Middle French period.

```
1 :appreper  a   ontolex:LexicalEntry , ontolex:Word ;
2 ontolex:canonicalForm :appreper_lemma .
3
4 :appreper_lemma  a   ontolex:Form ;
5 ontolex:writtenRep "appreper"@frm-x-00walloon .
```

In our language tag on Line 5, frm identifies it as from the Middle French period, with 00 indicating that it is a user-defined language (cp. E1).

3.3 Modern French

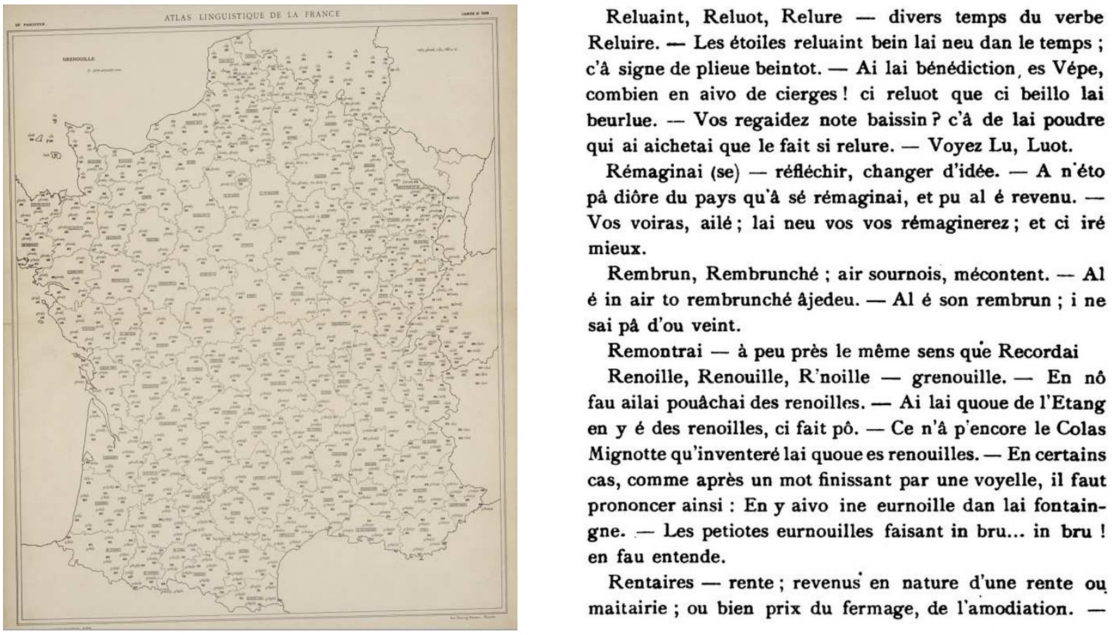
Today, standard French is dominant in all regions of France. Nevertheless, regional variation, dialects and patois characterize its linguistic landscape (Wolf, 1979: 165). This is illustrated, e.g., by the many dictionaries and surveys referenced by Lexilogos for French dialects. Attempts to revive regional varieties gave impetus to the creation of many linguistic atlases of France, beginning as early as 1897-1901 with the *Atlas linguistique de la France* – ALF (Gilliéron & Edmont, 1902–1910, Fig. 1a) and leading

[15] http://atilf.fr/dmf/definition/appreper [01-03-2019].

to the many large-sized volumes of the series *Atlas linguistiques de la France par régions* – ALFR (Séguy, 1973: 78).

The language code for Modern French is ISO 639-1 ‘fr’. For the majority of French regional varieties, ISO 639 codes are not available, exceptions being ISO 639-3 ‘nrf’ for the Norman dialect[16], ‘pcd’ for Picard, and ‘wln’ for Walloon.

Given the amount of linguistic resources with diatopic data for modern French, we have selected exemplary data, namely from dictionaries of different patois. We focus on one use case: the designations for the frog. To model the data simply using ‘fr’ as the language code does not account for the linguistic reality in the regions in our focus: it would render the diatopic variation generic. BCP 47 specifies a region sub-tag that is typically used to indicate (diatopic or diastratic) variation within a country or territory, the standard being a code from ISO 3166. However, ISO 3166 registers administrative (sub-)divisions (in our case, *régions* and *départements* of contemporary France) whose boundaries do not necessarily match the language boundaries.[17] Hence, we make use of the *privateuse* subtag and codes provided by Glottolog, e.g., for Burgundian in **E3** (‘bourg1247’), in line with the pattern in Table 1. However, the patois spoken in Burgundy (and in any other region) differ. It is thus necessary to further distinguish



Reluaint, Reluot, Relure — divers temps du verbe Reluire. — Les étoiles reluaint bein lai neu dan le temps ; c'â signe de plieue beintot. — Ai lai bénédiction, es Vépe, combien en aivo de cierges ! ci reluot que ci beillo lai beurlue. — Vos regaidez note baissin ? c'â de lai poudre qui ai aichetai que le fait si relure. — Voyez Lu, Luot.

Rémaginai (se) — réfléchir, changer d'idée. — A n'éto pâ diôre du pays qu'â sé rémaginai, et pu al é revenu. — Vos voiras, ailé ; lai neu vos vos rémaginerez ; et ci iré mieux.

Rembrun, Rembrunché ; air sournois, mécontent. — Al é in air to rembrunché âjedeu. — Al é son rembrun ; i ne sai pâ d'ou veint.

Remontrai — à peu près le même sens que Recordai

Renoille, Renouille, R'noille — grenouille. — En nô fau ailai pouâchai des renoilles. — Ai lai quoue de l'Etang en y é des renoilles, ci fait pô. — Ce n'â p'encore le Colas Mignotte qu'inventeré lai quoue es renouilles. — En certains cas, comme après un mot finissant par une voyelle, il faut prononcer ainsi : En y aivo ine eurnoille dan lai fontaingne. — Les petiotes eurnouilles faisant in bru... in bru ! en fau entende.

Rentaires — rente ; revenus en nature d'une rente ou maitairie ; ou bien prix du fermage, de l'amodiation. —

(a) (b)

Figure 1: (a) ALF map nº 668 ‘grenouille’. (b) Denizot (1910: 120).

[16] Falsely described as “Guernésiais, Jèrriais” which excludes the continental area.

[17] https://tools.ietf.org/html/bcp47#section-2.2.4; https://www.iso.org/obp/ui/#iso:code:3166:FR [1106-2019].

the language tag on patois. We do this by adding the name of the location where the patois has been recorded. This can be (1) a region or (2) a place name.

To identify a language in a region (1), as a subset of the language denoted by the Glottocode, we use the latitude and longitude coordinates of the location provided by the geographical database GeoNames[18] and we convert the coordinates into a Geohash[19], where Geohash is a system for encoding geographic coordinates as a base32 string, in a syntax acceptable for BCP 47 (Gillis-Webber & Tittel, 2019: 4:10). To identify a place name (2) within the language tag, we refer to its equivalent entry in GeoNames.

3.3.1 Language of Burgundy

E3, from *Dictionnaire de patois de Mancey* (Millot (1905–1922 (edition 1998)):

```
1   @PREFIX pwn:          <http://wordnet-rdf.princeton.edu/id/> .
2
3   :gornaille a    ontolex:LexicalEntry , ontolex:Word ;
4   :rdfs:label "gornaïlle"@fr-x-01bour1247-342996271 ;
5   ontolex:canonicalForm :gornaille_lemma ;
6   ontolex:sense        :gornaille_sense ;
7   ontolex:evokes       :frog_lexConcept.
8
9   :gornaille_lemma  a    ontolex:Form ;
10  ontolex:writtenRep       "gornaïlle"@fr-x-01bour1247-342996271 .
11
12  :gornaille_sense    a    ontolex:LexicalSense ;
13  ontolex:isLexicalizedSenseOf :frog_lexConcept .
14
15  :frog_lexConcept  a    ontolex:LexicalConcept ;
16  ontolex:lexicalizedSense :gornaille_sense ;
17  ontolex:isConceptOf      dbpedia:Frog ;
18  ontolex:definition       "grenouille"@fr ;
19  dct:references pwn:01642406-n .
```

In our language tag on Lines 4 and 10, fr identifies the tag as from the Modern French period, with 01 indicating that the Glottocode for the Burgundy language is used. To

[18] https://www.geonames.org/ [07-06-2019].

[19] https://www.movable-type.co.uk/scripts/geohash.html [07-06-2019].

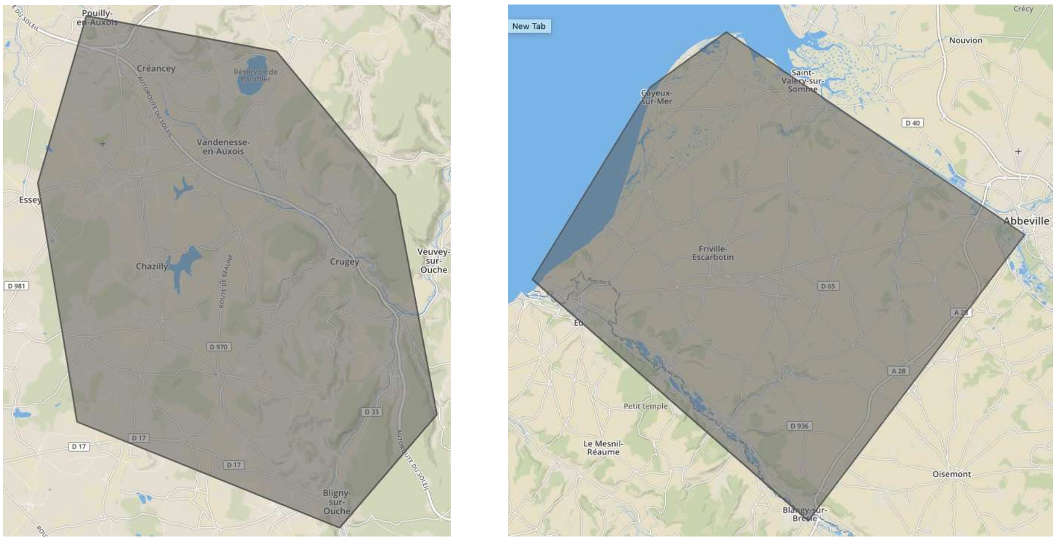
identify the patois spoken in Mancey, a commune in the Saône-et-Loire *département*, we made use of the equivalent identifier from GeoNames, 2996271, prepending it with 34 as per Table 1.

E4, from the *Vocabulaire patois de Sainte-Sabine et ses environs (Côte-d'Or)* (Denizot (1910), Fig. 1b):

```
1  :renoille   a    ontolex:LexicalEntry , ontolex:Word ;
2  rdfs:label  "renoille"@fr-x-00saintesabine-30u0g6r--
3  u0e36--u07zp--u0sbk--u0t5k--u0u4u ;
4  ontolex:canonicalForm :renoille_lemma ;
5  ontolex:sense  :renoille_sense ;
6  ontolex:evokes:frog_lexConcept .
7
8  :gueurnouille_lemma a ontolex:Form ;
9  ontolex:writtenRep       "renoille"@fr-x-00saintesabine-30u0g6r--
10 u0e36--u07zp--u0sbk--u0t5k--u0u4u .
11
12 :gueurnouille_sense a   ontolex:LexicalSense ;
13 ontolex:isLexicalizedSenseOf :frog_lexConcept .
```

The use of GeoNames to identify the location of Sainte-Sabine, a commune in the Côted'Or *département*, would be a wrong approach for this case: the title of the resource clearly indicates that the vocabulary has been recorded in Sainte-Sabine and, also, within its vicinity. Unfortunately, the introduction of the resource gives only a vague description of what it means: "montagnes des environs des Pouilly-en-Auxois et de Blignysur-Ouche", Denizot (1910: 14). We drew a polygon of the area that is, thus, only an approximation as well (Figure 2a). The geographic coordinates representing the polygon are: (49.62686,4.91473), (48.04287,4.66964), (47.6435,5.59192), (47.88325,6.85844), (48.40865,7.23867), (49.72584,5.81263), (49.62686,4.91473).

The last coordinate is the same as the first, and so we excluded the last one and then converted the latitude and longitude coordinates to a Geohash to a precision of five digits, *cf.* Gillis-Webber and Tittel (2019: 4:10f.): u0g6r--u0e36--u07zp--u0sbk--u0t5k--u0u4u. Lines 2-3 and 9-10 show the use of these Geohashes, with the pattern 00 defining the language as user-defined and 30 defining a geohashed polygon region.



(a) (b)

Figure2: (a) Approximate region where the patois of Sainte-Sabine was recorded. (b) Region of Vimeu in Picardy

3.3.2 Language of Picardy

E5, from Dictionnaire des parlers picards du Vimeu (Somme) (Vasseur (1998)):

```
1  :guernouille a    ontolex:LexicalEntry , ontolex:Word ;
2  rdfs:label
3  "guérnouille"@pcd-x-30u0cje--u0cj3--u0buz--u0chj--u0cm1 ;
4  ontolex:canonicalForm :guernouille_lemma ;
5  ontolex:sense     :guernouille_sense ;
6  ontolex:evokes    :frog_lexConcept .
7
8  :guernouille_lemma  a    ontolex:Form ;
9  ontolex:writtenRep
10 "guérnouille"@pcd-x-30u0cje--u0cj3--u0buz--u0chj--u0cm1 .
11
12 :guernouille_sense    a    ontolex:LexicalSense ;
13 ontolex:isLexicalizedSenseOf :frog_lexConcept .
```

In the language tag on Lines 3 and 10, the language code uses the ISO 639-3 code ‘pcd’ for the modern Picard language. To specify the region of Vimeu in Picardy (Fig. 2b), we have again defined a region, converted into Geohashes.

4. Modelling of regional variation using linguistic atlas data

We modeled a small set of exemplary data from the ALF. It seems clear to us that most of the regional differences manifested in a linguistic atlas concern phonetic variation. However, the regional particularities also concern the lexis, especially in border regions of France. These regions document phenomena of cultural and linguistic contact with other languages, e.g., with German, Franco-Provençal, Occitan, and Breton. These phenomena are of great interest, in particular to researchers in Historical Linguistics and Digital Humanities. With its rich lexical and phonetic data, an atlas could add significant value to the landscape of semantically accessible linguistic data sets.

For the transformation of linguistic atlas data into LD, the information on a map needs to be turned into points. This leads to two issues: dealing with (a) the geographic data acquisition points (which, in the context of ALF, is place names) and (b) the phonetic transcription indicated for each point.

For (a), Gally et al. (2013: 188f.) describe that they semi-automatically provided each of the 992 data acquisition points of the digitized ALF with geographic coordinates. For (b), typically, the data sources for the linguistic atlases are surveys where interviewees pronounced words and phrases and interviewers transcribed the phonetic realizations using a phonetic alphabet. For the ALF, Abbé Rousselot and Jules Gilliéron established a phonetic alphabet in 1891 which then was also used by the makers of the atlases of the series ALFR. The transcriptions were written onto the maps by hand. To ensure the structural interoperability of atlas data within the Semantic Web, the transcriptions need to be re-encoded using the standard *International phonetic alphabet* (IPA, International Phonetic Association, 2005), cp. Moran (2012) who uses IPA as an interlingual pivot for different transcription systems.

4.1 Exemplary data for Lorraine

We have used data from the ALF map n° 668 (Fig. 1a). In **E6**, for the lexeme *grenouille* "frog", we model the phonetic realizations of three acquisition points taken from the Meurthe-et-Moselle *département* in Lorraine (Table 2) using the phoneticRep property of the OntoLex-Lemon vocabulary.

- n° 162 (Sexey-les-Bois)
- n° 170 (Moncel-sur-Seille)
- n° 171 (Mailly-sur-Seille)

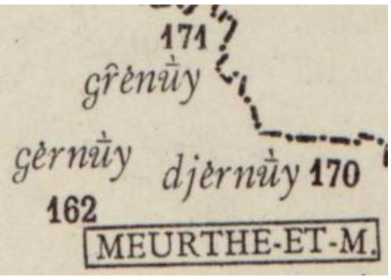


Table 2: Extract from ALF map nº 668.

E6, from *Atlas linguistique de la France* (Gilliéron & Edmont, 1902–1910):

```
1   :grenouille   a     ontolex:LexicalEntry , ontolex:Word ;
2   rdfs:label          "grenouille"@fr ;
3   ontolex:canonicalForm :grenouille_lemma ,
4   ontolex:sense       :grenouille_sense ;
5   ontolex:evokes      :frog_lexConcept .
6
7   :grenouille_lemma   a    ontolex:Form ;
8   ontolex:writtenRep "grenouille"@fr ;
9   ontolex:phoneticRep "gK@nu–:j"@fr-fonipa-x-01lorr1242-342996683 ,
10                      "g@rnu–:j"@fr-fonipa-x-01lorr1242-342974669 ,
11                      "dZ@rnu–:j"@fr-fonipa-x-01lorr1242-342993415 .
12
13   :grenouille_sense     a ontolex:LexicalSense ;
14  ontolex:isLexicalizedSenseOf :frog_lexConcept .
```

In Lines 9-11, we have re-encoded the phonetic transcription (*cf.* Table 2) using IPA characters. To identify the phonetic characters of the string literals, we include the subtag fonipa, which is compliant with BCP 47 (Phillips & Davis, 2009: 43). In the *privateuse* portion, 01 indicates a code from Glottolog has been used. As with **E3**, the place name for each geographic acquisition point has been represented by its equivalent GeoNames identifier, prepended with 34. E.g., the phonetic representation of the lexeme recorded in Sexey-les-Bois (nº 162, Line 10) is identified as 2974669.[20]

[20] http://www.geonames.org/2974669/sexey-les-bois.html [06-06-2019].

5. Discussion

Revisiting the CQs, all questions, with the exception of **CQ6**, are answerable with the available data from our case study.

CQ1 is answered by **E1**–**E4** and **E6**. For **E1** and **E2**, codes exist in alternative directories, but they do not reflect the correct time periods. Hence, we opted to identify the language using a user-defined code, indicated by 00 from Table 1. **CQ2** is, thus, also answered by these two exemplars. For **E3**, **E4** and **E6**, a Glottocode is available, indicated by 01 from Table 1.

CQ3 is answered by our Modern French exemplars. Although different language codes are available for Modern French in each ISO 639 part, we make use of 'fr' from ISO 639-1; as per the BCP 47 specification, the shortest language code available has to be used.

CQ4 is answered by **E3**–**E6** showing two solutions: (1) **E3** and **E6** make use of an identifier from GeoNames, indicated by 34 from Table 1, (2) **E4** and **E5** both make use of a user-defined language (defined with pattern 00) and of Geohashes that represent the geographic coordinates for a polygon shaped region (defined with pattern 30 and with -- serving as an internal delimiter between each Geohash). A detailed description of associating a geographic area with a language is discussed in Gillis-Webber and Tittel (2019), which also addresses **CQ5**.

Although the pattern allows for a more precise definition of the language in question, for **E4** and **E5** the language tags intuitively feel too long: the Geohash, while useful, is opaque, and may require further annotation in order to be human-readable. While the proposed pattern serves as an interim solution for language-tagging lesser-known or less-discussed languages, the problem still remains that the dependency of a language tag on an ISO standard or registry is a flaw of language tags and the RDF specification. As an alternative to a language tag, we should be able to encode a URI in the vein of "jannaie"@deaf:fro/gallo, where deaf: is the namespace.

Gillis-Webber and Tittel (2019) suggest exploring the creation of a sub-datatype for rdf:langString, which would thus allow for the datatype URI to be encoded, as an alternative to the language tag. However, doing this presents challenges. A literal consists of two elements: a lexical form and a datatype URI (Cyganiak et al., 2014). If the datatype URI is http://www.w3.org/1999/02/22rdf-syntax-ns#langString, then a third element is introduced to the literal: namely "a non-empty language tag as defined by BCP 47", *ib.* All other datatype URIs are mapped to RDF-compatible XSD types, none of which would allow the introduction of a custom URI in the place of a language tag, *ib.* To allow for an alternative datatype URI, the RDF specification would have to be amended. However, as a sub-datatype of rdf:langString, the constraints of BCP 47 would still apply. It thus seems easier to propose a change to BCP 47: namely to allow, for the *privateuse* sub-tag only, the following characters: [-:/a-zA-Z0-9]. This would then render a language tag of the form "jannaie"@x-deaf:fro/gallo. To be RDF-compatible, the namespace for x-deaf: would have to be defined in the same RDF document in which the language tag is used.

We considered creating a user-defined simple XML Schema datatype, as a restriction on an existing datatype (Carroll & Pan, 2006). Although it would not render a language tagged string literal, it would render a string literal with an encoded URI: "jannaie"^^ <http://example.org/simpleTypes#froGallo>. However, the URI, although it clearly identifies the language, would not be dereferenceable which is in opposition to one of the principles of LD. Furthermore, it would not be appropriate for use when modelling data using Ontolex-Lemon because the latter requires rdf:langString when representing forms. This leads us to conclude that Part 4 is required in our pattern, i.e., for the inclusion of a URI shortcode in the *privateuse* portion of a language tag, which can then be mapped to a URI.

Apart from the question of how to design the language tags, a further question arises: is the granularity of our approach sufficient for the following scenarios? The language of a linguistic resource, e.g., a text or a dictionary, is written:

1. during a time span or covering a time span, e.g., a collection of 19th century legal documents or a dictionary covering several centuries such as the DEAF,

2. at different times, e.g., the *Roman de la Rose* that consists of two parts (ca.1230; ca.1275) by two authors[21],

3. in different places or covers several places, some parts (in) region A, some parts (in) Region B.

The scenarios describe multilingual settings that require multilingual labels (a part of the RDF standard[22]). Scenarios 1 and 2 can be answered with the range of Part 2 of our pattern. For scenario 3, two questions arise: how to identify (a) the language(s) of a triple subject (a lexicon, a lexical entry, etc.), and (b) the language(s) of a literal.

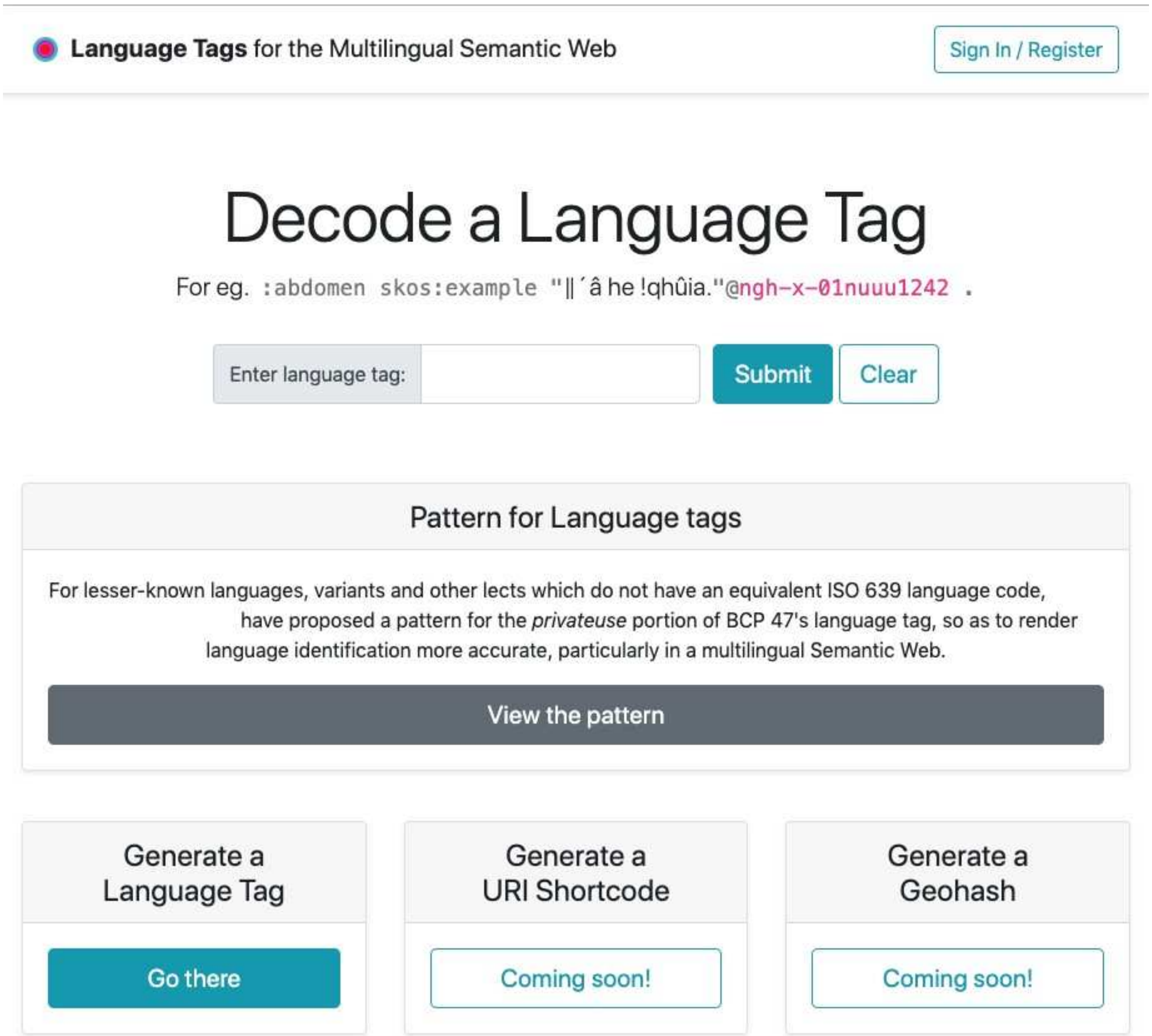
Question (a) is answerable with the property dct:language that has multiple values, such as <http://example.org/language-1> and <http://example.org/language-2> respectively (cp. **E0** with both ISO 639-1 and ISO 639-3 code). Question (b) is answerable with multiple literals, i.e., duplicated language-tagged literals for the same subject and predicate, with a custom language tag for each.

[21] http://www.deaf-page.de/bibl/bib99r.php#RoselLangl [11-06-2019].

[22] https://www.w3.org/community/bpmlod/wiki/Best_practises_-_previous_notes [12-06-2019].

6. Interface for Language Tag Generation

A user interface and REST API to both generate and decode language tags, currently in development, is to be demonstrated at eLex 2019. Language tags can be generated according to our pattern. For the decoding of language tags, the results are available in JSON, with natural language, RDF/XML and Turtle syntax to follow. Figure 3 shows the user interface. See https://londisizwe.org/language-tags/ for more information.



For eg. `:abdomen skos:example "ǁ´â he !qhûia."@ngh-x-01nuuu1242 .`

For lesser-known languages, variants and other lects which do not have an equivalent ISO 639 language code, have proposed a pattern for the *privateuse* portion of BCP 47's language tag, so as to render language identification more accurate, particularly in a multilingual Semantic Web.

Figure 3: User interface for generating and decoding language tags.

7. Conclusions & Future Work

In this paper, we have discussed how to create language tags when modelling linguistic data as LD for languages for which ISO 639 does not provide language codes. We have focused on linguistic resources of French that are of interest for diatopic studies, and we have chosen exemplary data with a diachronic view, including Old-, Middle- and Modern French dictionaries and a Modern French linguistic atlas. For each exemplar, we have created a language tag, in line with a proposed pattern. These language tags identify the language, its historical language stage, a subset of the language (dialect or patois) in an unambiguous way. Using a URI shortcode, the language tags can be reduced to a more user-friendly length. This, however, makes them opaque, whereas the former is more descriptive but can be long. While the use of encoded URIs affects human-readability, it remains machine-readable nonetheless.

Extension towards MoLA. In collaboration with C. Maria Keet, the authors have been working on MoLA, a Model for Language Annotation (Gillis-Webber et al., 2019). MoLA is a lightweight ontology which allows for languoids (a language family, language, dialect cluster, or lect) to be represented in RDF. Due to its expressiveness, including MoLA in the modelling of linguistic resources enables comprehensive language information to be represented. Future work is, thus, to model the languages identified in these French resources using MoLA.

Other Resources. We conclude the paper returning to linguistic desiderata: Other linguistic atlases (of the series ALFR, e.g., Lanher et al., 1979–1988 [Lorraine Romane]; Dondaine & Dondaine, 1972–1991 [Franche-Comté]) and dictionaries should be evaluated for a future conversion to LD. Valuable dictionaries comprise those covering particular patois and dialects, the comprehensive dictionary of French regionalisms (Rézeau, 2001), etc. The modelling of lexicologically rich resources of other kinds is a further task, including a lexicographer's standard work for historic botany, the *Flore populaire de la France…* (Rolland, 1896–1914), and corpora, e.g., the *Corpus Historique du Substandard Français* (CHSF, Thun, 2011).

8. Varieties of French

Figures 4 and 5 show the designations of French varieties, the corresponding Glottocodes and ISO 639-3 codes, respectively. We define the lists of Old French varieties given by the FEW (von Wartburg (1922–: *Beiheft* p.63)) and by the DEAF as authority lists and exclude all regional varieties listed by other resources (e.g., Lexilogos) that are not covered by the FEW- or the DEAF list.

Modern French / FEW	Old French / FEW	Old French / DEAF	Glottolog (modern)	ISO 639-3 (modern)
français moderne	—	français moderne	stan1290	fra
—	ancien français	ancien français	—	fro *
—	moyen français	moyen français	mid1316	frm *
—	—	francien	—	—
pik.	apik.	picard	pica1241 **	pcd
hain.	—	hennuyer	hain1252	—
art.	—	artésien	arto1238	—

wallon	awallon.	wallon	wall1255	wln
lütt.	alütt.	liégeois	—	—
nam.	anam.	—	—	—
flandr.	aflandr.	français de la Flandre française	—	—
Lille	alill.	—	lill1247	—
champ.	achamp.	champenois	—	—
lothr.	alothr.	lorrain	lorr1242	—
norm.	anorm.	normand	norm1245	nrf
—	agn.	anglo-normand	angl1258	xno *
hbret.	—	haut-breton	gall1275	—

* Historical language stage. ** 12 sub-languages incl. ‘hain1252’, ‘arto1238’, ‘lill1247’.

Figure 4: List of French varieties, part 1 (terms in French).

Modern French / FEW	Old French / FEW	Old French / DEAF	Glottolog (modern)	ISO 639-3 (modern)
ang.	—	angevin	ange1244	—
poit.	apoit.	poitevin	poit1240	—
saint.	—	saintongeais	sant1407	—
tour.	—	tourangeau	—	—
orl.	—	orléanais	—	—
bourbonn.	abourb.	bourbonnais	bour1246	—
bourg.	abourg.	bourguignon	bour1247	—
Lyon **	—	lyonnais	lyon1243 ***	—
frcomt.	afrcomt.	franc-comtois	fran1262 ***	—
—	—	franco-italien	—	—
—	—	Nord-Est	—	—
—	—	Nord	—	—
—	—	Nord-Ouest	—	—
—	—	Ouest	—	—
—	—	Sud-Ouest	—	—
centr.	—	Centre	—	—
—	—	Est	—	—
—	—	Sud-Est	—	—
—	—	Terre Sainte	—	—
—	judfr.	Judeofrançais	—	zrp *

* Historical language stage. ** Sub Savoy. *** Sub Francoprovençalic.

Figure 5: List of French varieties, part 2 (terms in French).

9. References

ATILF – CNRS & Université de Lorraine (2015). *Dictionnaire du Moyen Français, version 2015 (DMF 2015).* Paris. URL http://www.atilf.fr/dmf/. Accessed: 17-06-2019.

Baldinger, K., Möhren, F. & Städtler, T. (1971–). *Dictionnaire étymologique de l'ancien français (DEAF).* Québec / Tübingen / Berlin: Presses de L'Université Laval / Niemeyer / De Gruyter. DEAF*él*: https://deaf-server.adw.uni-heidelberg.de].

Berners-Lee, T. (2006). *Linked Data.* World Wide Web Consortium. URL https://www.w3.org/DesignIssues/LinkedData.html. Accessed: 17-06-2019.

Berschin, H., Felixberger, J. & Goebl, H. (2008). *Französische Sprachgeschichte.* Hildesheim / Zürich / New York: Olms.

Bizer, C., Heath, T. & Berners-Lee, T. (2009). Linked Data – The Story So Far. *International Journal on Semantic Web and Information Systems*, 5, pp. 1–22.

Bouda, P. & Cysouw, M. (2012). Treating Dictionaries as a Linked-Data Corpus. In C. Chiarcos (ed.) *Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata.* Berlin/Heidelberg, Germany: Springer, pp. 15–23.

Carroll, J. & Pan, J. (2006). XML schema datatypes in RDF and OWL: W3C Working Group Note 14 March 2006. URL https://www.w3.org/TR/swbp-xsch-datatypes/. Accessed: 17-06-2019.

Chambon, J. P. (1997). Pour la localisation d'un texte de moyen français: le Mystère de Saint Sébastien. In G. Kleiber & M. Riebel (eds.) *Les formes du sens: Etudes de linguistique française, médiévale et générale offertes à Robert Martin à l'occasion de ses 60 ans.* Louvain-la-Neuve: Duculot, pp. 201–216.

Chauveau, J. P. (2016). Régionalismes médiévaux et dialectismes contemporains en hauteBretagne. In M. Glessgen & D. Trotter (eds.) *La régionalité lexicale du français au Moyen Âge.* Strasbourg: ÉLiPhi, pp. 131–166.

Chiarcos, C., McCrae, J., Cimiano, P. & Fellbaum, C. (2013). Towards Open Data for Linguistics: Lexical Linked Data. In A. Oltramari, P. Vossen & L. Qin et al. (eds.) *New Trends of Research in Ontologies and Lexical Resources: Ideas, Projects, Systems.* Berlin / Heidelberg: Springer, pp. 7–25.

Cyganiak, R., Wood, D. & Lanthaler, M. (2014). RDF 1.1. concepts and abstract syntax: W3C recommendation 25 February 2014. URL https://www.w3.org/TR/2014/ REC-rdf11-concepts-20140225/. Accessed: 17-06-2019. de Melo, G. (2015). Lexvo.org: Language-Related Information for the Linguistic Linked Data Cloud. *Semantic Web*, 6(4), pp. 393–400.

Denizot, J. (1910). *Vocabulaire patois de Sainte-Sabine et ses environs (Côte-d'Or).* Beaune: Imprimerie Beaunoise.

Dondaine, C. & Dondaine, L. (1972–1991). *Atlas linguistique et ethnographique de la Franche-Comté (ALFC).* Paris: Éd. du CNRS.

Gally, S., Chauvin-Payan, C. & Davoine P. A. et al. (2013). GéoDialect : Exploration des outils géomatiques pour le traitement et l'analyse des données géolinguistiques. *Géolinguistique*, 14, pp. 186–208.

Gillis-Webber, F. & Tittel, S. (2019). The Shortcomings of Language Tags for Linked Data when Modeling Lesser-Known Languages. In *Proceedings of LDK2019, Leipzig, Germany, 21-22 May 2019, OASIcs, Vol. 70.* pp. 4:1–4:15.

Gillis-Webber, F., Tittel, S. & Keet, M. (2019). A Model for Language Annotations on the Web. In B. Villazón-Terrazas & Y. Hidalgo-Delgado (eds.) *Knowledge Graphs and Semantic Web. 1st Iberoamerican Conference, KGSWC 2019, Villa Clara, Cuba, June 23-30, 2019, Proceedings.* pp. 1–16.

Gilliéron, J. & Edmont, E. (1902–1910). *Atlas linguistique de la France.* Paris: Champion. Glessgen, M. & Trotter, D. (2016). *La régionalité lexicale du français au Moyen Âge.* Strasbourg: ÉLiPhi.

Gleßgen, M. D. & Thibaut, A. (2005). La «régionalité linguistique»: essai définitoire. In M.D. Gleßgen & A. Thibaut (eds.) *La lexicographie différentielle du français et le Dictionnaire des régionalismes de France.* Presses Univ. de Strasbourg, pp. III–XVII.

International Organization for Standardization (n.d.). Language codes – ISO 639. URL https://www.iso.org/iso-639-language-codes.html. Accessed: 17-02-2019.

International Phonetic Association (2005). International Phonetic Alphabet. Tech. rep. URL https://www.internationalphoneticassociation.org/. Accessed: 17-02-2019.

Lanher, J., Litaize, A. & Richard, J. (1979–1988). *Atlas linguistique et ethnographique de la Lorraine Romane (ALLR).* Paris: Éd. du CNRS.

Millot, C. (1905–1922 (edition 1998)). *Dictionnaire de patois de Mancey.* Tournus: Société des amis des arts et des sciences de Tournus.

Moran, S. (2012). Using Linked Data to Create a Typological Knowledge Base. In C. Chiarcos (ed.) *Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata.* Springer, pp. 129–138.

Möhren, F. (2016). La régionalité dans le DEAF – historique et programme. In M. Glessgen & D. Trotter (eds.) *La régionalité lexicale du français au Moyen Âge.* Strasbourg: ÉLiPhi, pp. 37–50.

Phillips, A. & Davis, M. (2009). Tags for Identifiying Languages. *BCP*, 47. URL https://tools.ietf.org/html/bcp47. Accessed: 17-06-2019.

Renders, P. (2016). La régionalité lexicale du moyen français (1350–1500). In M. Glessgen & D. Trotter (eds.) *La régionalité lexicale du français au Moyen Âge.* Strasbourg: ÉLiPhi, pp. 85–96.

Rickard, P. (1974). *A history of the French language.* London: Hutchinson University Library.

Rolland, E. (1896–1914). Flore populaire de la France ou histoire naturelle des plantes dans leurs rapports avec la linguistique et le folklore. Paris: Rolland.

Rézeau, P. (ed.) (2001). *Dictionnaire des régionalismes de France. Géographie et histoire d'un patrimoine linguistique.* Bruxelles: De Boeck.

Rézeau, P. (2007). *Richesses du français et géographie linguistique.* Bruxelles: De Boeck & Larcier.

Rézeau, P. (2016). La régionalité lexicale du français après 1500, à travers des régionalismes recueillis dans les correspondances de poilus. In M. Glessgen & D. Trotter (eds.) *La régionalité lexicale du français au Moyen Âge.* Strasbourg: ÉLiPhi, pp. 111–130.

Séguy, J. (1973). Les Atlas linguistiques de la France par régions. *Langue Française*, 18, pp. 65–90.

Thun, H. (2011). Die diachrone Erforschung der *français régionaux* auf der Grundlage des *Corpus Historique du Substandard Français.* In C. Schlaak & L. Busse (eds.) *Sprachkontakte, Sprachvariation und Sprachwandel.* Narr, pp. 359–394.

Tittel, S. (2016). La régionalité lexicale de l'ancien français (ca.1100 – ca.1350) : Une enquête sur la base du *Dictionnaire étymologique de l'ancien français.* In M. Glessgen & D. Trotter (eds.) *La régionalité lexicale du français au Moyen Âge.* Strasbourg: ÉLiPhi, pp. 61–84.

Tittel, S., Bermúdez-Sabel, H. & Chiarcos, C. (2018). Using RDFa to Link Text and Dictionary Data for Medieval French. In J. P. McCrae, C. Chiarcos & T. Declerck et al. (eds.) *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). 6th Workshop on Linked Data in Linguistics (LDL-2018), 12 May 2018, Miyazaki, Japan.* Paris: ELRA, pp. 30–38.

Tittel, S. & Chiarcos, C. (2018). Historical Lexicography of Old French and Linked Open Data: Transforming the Resources of the *Dictionnaire étymologique de l'ancien français* with OntoLex-Lemon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). GLOBALEX Workshop (GLOBALEX-2018), Miyazaki, Japan, 2018.* Paris: ELRA, pp. 58–66.

Vannini, L. & Le Crosnier, H. (2012). *Net.lang. Towards the multilingual cyberspace.* Caen: C & F Éditions.

Varlet, M. (1896). *Dictionnaire du patois meusien.* Verdun: Société Philomathique de Verdun.

Vasseur, G. (1998). *Dictionnaire des parlers picards du Vimeu (Somme), avec index français-picard.* Fontenay-sous-Bois: SIDES.

von Wartburg, W. (1922–). *Französisches Etymologisches Wörterbuch (FEW).* Bonn, Heidelberg, Leipzig/Berlin, Basel: ATILF. [Continued by O. Jänicke, C. T. Gossen, J. P. Chambon, J.-P. Chauveau, and Yan Greub].

Wolf, H.J. (1979). *Französische Sprachgeschichte.* Heidelberg: Quelle u. Meyer.