UNIVERSITY OF CAPE TOWN

The construction of a linguistic linked data framework for bilingual lexicographic resources

Author:

Frances Gillis-Webber

Supervisors:

Richard Higgs, Connie Bitso

A minor dissertation submitted in partial fulfilment of the requirements for the award of the degree of Master of Philosophy, specialisation in Digital Curation.

Department of Knowledge and Information Stewardship

2018

(Intentionally left blank for copyright notice)

Plagiarism Declaration

I understand the meaning of plagiarism and declare that all of the work in the document, save for that which is properly acknowledged, is my own.

Signed by candidate

Frances Gillis-Webber

26 November 2018

Acknowledgements

I would like to express my deep gratitude to my supervisor, Richard Higgs; he has been a central figure throughout my degree, and this dissertation is the culmination of his commitment, support and depth of knowledge, of which he freely gave, providing succour when I needed it most.

I would also like to thank the Head of Department, Associate Professor Jayarani Raju, for her kindness and ongoing support. The Department of Knowledge and Information Stewardship staff complement may be small but its dedication to its students far exceeds expectations, for which they can be immensely proud.

A big thank you to the library and the benevolence of the librarians at the front desk: many a fine was waived and my book quota often allowed to be exceeded. A special thanks to Rosie and Russell at InterLibrary Loans for sourcing so many books and papers – without InterLibrary Loans this research would not have been possible.

. . . . .

On a personal note, this dissertation is dedicated to my late mother, Gillian Lesley Webber ... so many dreams.

Abstract

The construction of a linguistic linked data framework for bilingual lexicographic resources

by Frances Gillis-Webber

Little-known lexicographic resources can be of tremendous value to users once digitised. By extending the digitisation efforts for a lexicographic resource, converting the human-readable digital object to a state that is also machine-readable, structured data can be created that is semantically interoperable, thereby enabling the lexicographic resource to access, and be accessed by, other semantically interoperable resources.

The purpose of this study is to formulate a process when converting a lexicographic resource in print form to a machine-readable bilingual lexicographic resource applying linguistic linked data principles, using the *English-Xhosa Dictionary for Nurses* as a case study. This is accomplished by creating a linked data framework, in which data are expressed in the form of RDF triples and URIs, in a manner which allows for extensibility to a multilingual resource. Click languages with characters not typically represented by the Roman alphabet are also considered. The purpose of this linked data framework is to define each lexical entry as "historically dynamic", instead of "ontologically static" (Rafferty, 2016:5). For a framework which has instances in constant evolution, focus is thus given to the management of provenance and linked data generation thereof. The output is an implementation framework which provides methodological guidelines for similar language resources in the interdisciplinary field of Library and Information Science.

Keywords: lexicography, bilingualism, linguistics, ontology, linked data, implementation framework

(Intentionally left blank)

Table of Contents

List of Figures

List of Tables

List of Code Examples

List of Acronyms & Abbreviations

BCP	Best Current Practice
BFO	Basic Formal Ontology
CC	Creative Commons
CILI	Collaborative Interlingual Index
DCMI	Dublin Core Metadata Initiative
DOLCE	Descriptive Ontology for Linguistic and Cognitive Engineering
DWS	Dictionary Writing System
HLT	Human Language Technology
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
IRI	Internationalized Resource Identifier
ISO	International Organisation for Standardization
KOS	Knowledge Organisation System
L1	First language (when referring to language acquisition)
L2	Second language (when referring to language acquisition)
LIR	Linguistic Information Repository
LIS	Library and Information Science
LL(O)D	Linguistic Linked (Open) Data
LOD	Linked Open Data
LMF	Lexical Markup Framework
LR	Language Resource
MIME	Multipurpose Internet Mail Extensions
MSW	Multilingual Semantic Web
NID	National Institute for the Deaf
NLP	Natural Language Processing
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OWL	Web Ontology Language
PanSALB	Pan South African Language Board
PHP	Hypertext Preprocessor
POS	Part of Speech
PWN	Princeton WordNet
RDF	Resource Description Framework
RDFS	Resource Description Framework Schema
SASL	South African Sign Language
SKOS	Simple Knowledge Organisation System
SUMO	Suggested Upper Merged Ontology
UML	Unified Modelling Language
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
W3C	World Wide Web Consortium
XML	eXtensible Markup Language

(Intentionally left blank)

Chapter 1 **Introduction**

1.1 Introduction

A dictionary is a lexicographic resource which has considerable utility value, with each resource created to serve the specific needs of the intended target user. A dictionary can take various forms: from linguistic, focussing on the linguistic and pragmatic aspects of language, to encyclopaedic, providing extra-linguistic aspects as well (Gouws & Prinsloo, 2005:48; Zgusta, 1971:198). It can be a pedagogical dictionary, presented as a children's picture dictionary with a simplified vocabulary of core terms, or a specialised dictionary of technical terms; and if a linguistic dictionary, it can be monolingual or bilingual, providing translation equivalents for two or more vocabularies (Gouws & Prinsloo, 2005:48; Zgusta, 1971:198). A dictionary can be prescriptive, describing how a language should be used, or standard, describing how a language *is* used (Gouws & Prinsloo, 2005:2). The dictionary can also be diachronic, describing the vocabulary as it changes over time; or synchronic, with the vocabulary provided representing a period in time for the language (Zgusta, 1971:200-203). Whatever their form, lexicographic resources can be of tremendous value as a language resource to users once digitised, not least by rendering them more accessible, particularly lesser-known resources (European Commission, 2011; Chowdhury, 2015:43-44; McArthur, 1998:26). By extending the digitisation efforts for a lexicographic resource, converting the human-readable digital object to a state that is also machine-readable, structured data can be created that is semantically interoperable, thereby enabling the lexicographic resource to access, and be accessed by, other semantically interoperable resources (Arp, Smith & Spear, 2015:38).

This dissertation explores the construction of a framework for bilingual lexicographic resources, applying linguistic linked data principles. The methodology for converting a lexicon derived from a printed dictionary to structured data published on the web – in the form of Linked Data – is detailed, resulting in a framework which provides guidelines for similar applications.

1.2 Background to this study

Digitisation is the process of converting analogue resources into digital resources, typically from paper to image (Chowdhury, 2015:37). Retrodigitisation is the process of converting digital resources from simple digital objects (for example, an image in JPEG format) into complex digital objects with a machine-readable format (Raghallaigh & Měchura, 2014:67; Higgins, 2016:33-34). The form of these complex digital objects can vary; examples include: a collection of HTML files with semantic markup which describes the content therein; the same content but in XML format; a dataset stored in a relational database, or a dataset stored as Resource Description Framework (RDF) triples.

For the latter example, RDF is a specification by the World Wide Web Consortium (W3C) and it is a very simple data model consisting of triples, with each fact described as a short statement comprising a subject, predicate and an object (Van Hooland & Verborgh, 2014:3). Examples of projects built on the RDF data model are DBpedia[1], a knowledge base which extracts structured information from Wikipedia, and BabelNet[2], a multilingual encyclopaedic dictionary. WordNet[3], a lexical database for English, developed and maintained by Princeton University, is also supplied in RDF ("WordNet RDF", n.d.).

An example of two lexical entries, *abdomen* and *isisu*, expressed in short statements would be:

> *Abdomen* is a lexical entry.
> *Abdomen* is a word.
> *Abdomen* is a noun.
> *Abdomen* is an English term.
>
> *Isisu* is a lexical entry.
> *Isisu* is a word.

[1] http://wiki.dbpedia.org/
[2] http://babelnet.org/
[3] http://wordnet-rdf.princeton.edu/

Isisu is a noun.

Isisu is a Xhosa term.

Isisu is the equivalent of *Abdomen*.

There are different ways to write RDF triples (called 'serialisation'); common serialisation formats include Turtle, N-Triples and JSON-LD (World Wide Web Consortium [W3C], 2014e) (see Appendix B for further elaboration and code samples). A human-readable form of RDF serialisation, Turtle, is shown below for the same short statements:

Subject	*Predicate*	*Object*
:abdomen	isA	:lexicalEntry ;
	isA	:word ;
	isA	:noun ;
	isLanguage	:English .
:isisu	isA	:lexicalEntry ;
	isA	:word ;
	isA	:noun ;
	isLanguage	:Xhosa ;
	isEquivalent	:abdomen .

When these triples are visualised using node-edge-node structure as shown in Figure 1-1, the relationships between the two lexical entries become clearer (Van Hooland & Verborgh, 2014:3).

Tim Berners-Lee, inventor of the World Wide Web, said in a 2009 TED talk "data is about relationships", and by constructing relationships in the data, "the more things you have to connect together, the more powerful it is" (Berners-Lee, 2009). To construct these relationships, he suggests putting data on the web and to use Uniform Resource Identifiers (URIs). Although using the same Hypertext Transfer Protocol (HTTP) as Uniform Resource Locators (URLs), URIs differ from URLs conceptually in that they do not refer to the location of a document, instead, they identify: not just documents, but

any kind of object or concept (Berners-Lee, 2006; Berners-Lee, 2009; Hitzler, Krötzsch & Rudolph, 2010:21-22; Hyvönen, 2012:25; Van Hooland & Verborgh, 2014:46).

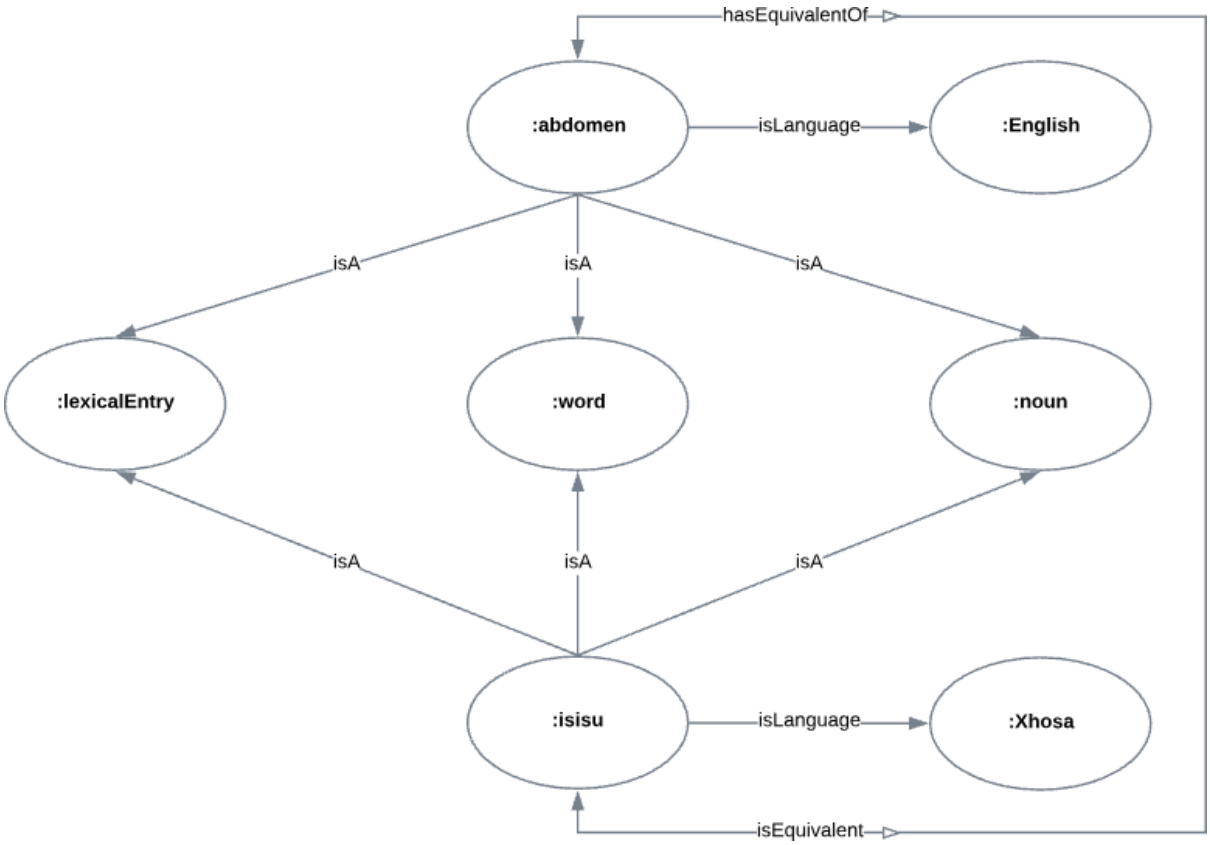


Figure 1-1: Visualisation of two lexical entries

Continuing with the example of the lexical entries, the :English and :Xhosa objects in the triples can be replaced with URIs from an external data source. Lexvo.org, an ontology for language-related entities, has been selected for this purpose. An additional triple has been added, where the lexical entry *abdomen* is identified to be denoted by the DBpedia resource "Abdomen". Because *isisu* is equivalent to *abdomen*, the relationship between *isisu* and the DBpedia resource can be inferred. The triples are now presented as:

Subject	*Predicate*	*Object*
:abdomen	isA	:lexicalEntry ;

	isA	:word ;
	isA	:noun ;
	isLanguage	http://lexvo.org/id/iso639-3/eng ;
	isDenotedBy	http://dbpedia.org/resource/Abdomen .
:isisu	isA	:lexicalEntry ;
	isA	:word ;
	isA	:noun ;
	isLanguage	http://lexvo.org/id/iso639-3/xho ;
	isEquivalent	:abdomen .

These triples are again visualised using node-edge-node structure in Figure 1-2.

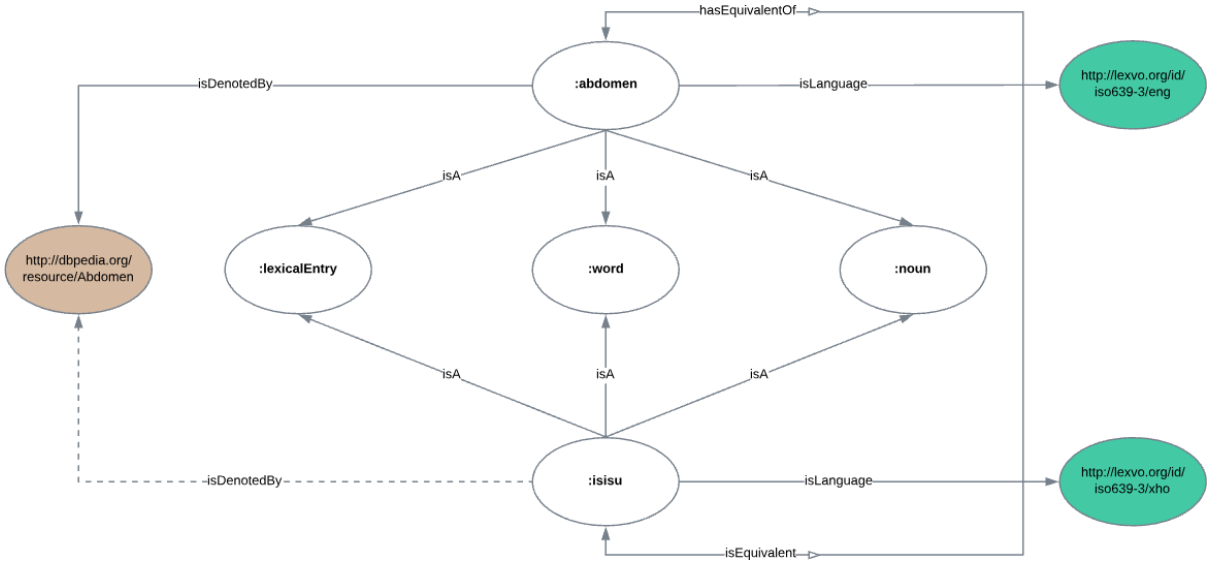


Figure 1-2: Adding URIs from external data sources

As shown in Figure 1-2, each lexical entry is linked to an external data source. Conversely, external data sources should be able to link to the lexical entries, also using URIs. This is done by defining a namespace for the dataset, for example: https://londisizwe.org/entry/, and setting a unique identifier for each lexical entry (shown in Figure 1-3).

The triples would then change to:

Subject *Predicate* *Object*

https://londisizwe.org/entry/en-n-abdomen

isA	:lexicalEntry ;
isA	:word ;
isA	:noun ;
isLanguage	http://lexvo.org/id/iso639-3/eng ;
isDenotedBy	http://dbpedia.org/resource/Abdomen .

https://londisizwe.org/entry/xh-n-isisu

isA	:lexicalEntry ;
isA	:word ;
isA	:noun ;
isLanguage	http://lexvo.org/id/iso639-3/xho ;
isEquivalent	:abdomen .

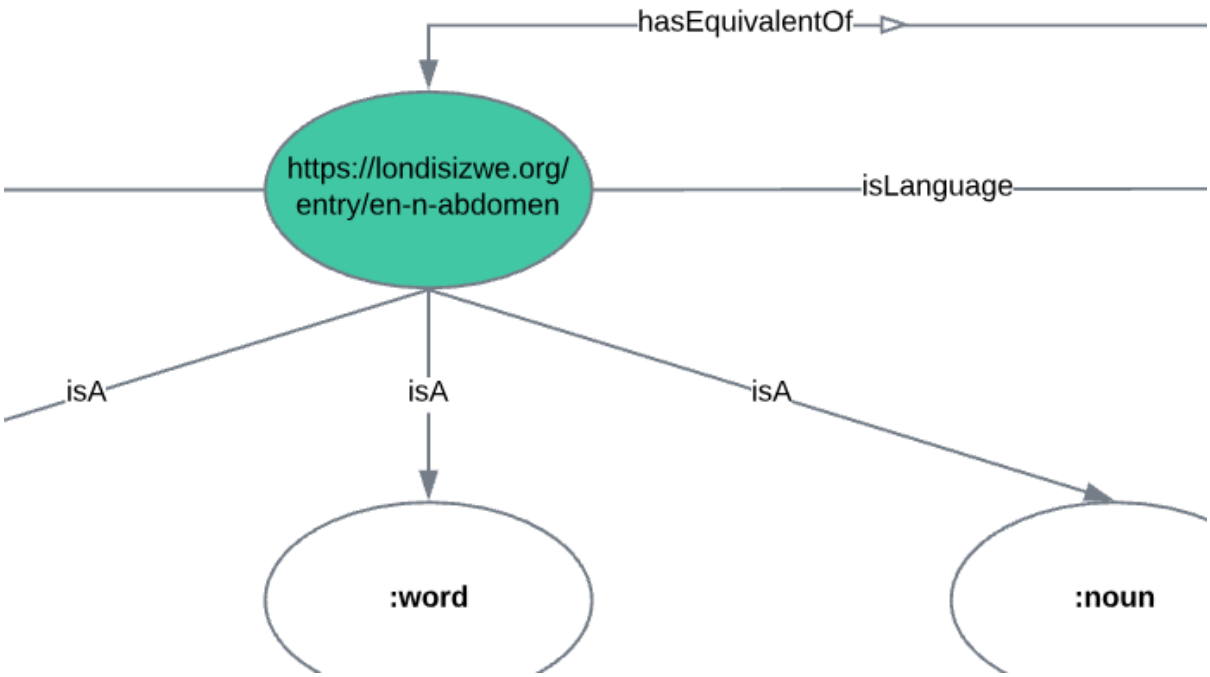


Figure 1-3: Converting the subject to a URI

Each subject, predicate or object, unless it has a literal value, can be converted to a URI, either by using a URI from an external data source, or by creating a URI using a predefined namespace. Identifying and creating links between data elements using URIs is a fundamental component when publishing structured data on the web, and the set of techniques and best practices for doing so is referred to as Linked Data (Wood et al., 2014:4; Van Hooland & Verborgh, 2014:3). The principles of Linked Data, applied to the bilingual lexicographic resource, *English-Xhosa Dictionary for Nurses* (EXDN), will form the basis for the remaining chapters of this study.

1.2.1 English-Xhosa Dictionary for Nurses

The artefact is the second edition of a bilingual, specialised dictionary of medical terms, compiled by Neil MacVicar, a medical doctor, in collaboration with isiXhosa-speaking nurses, published by Lovedale Press, a South African publisher, in 1935 (see Appendix A). The researcher concluded that if the text is no longer under copyright, it would be beneficial to digitise it; however, digitisation is only intended to render an artefact human-readable, not machine-readable as well. The researcher determined that using this little-known lexicographic resource as a case study would be suitable as a proof of concept for extending the digitisation efforts of a lexicographic resource into machine readability. By converting the human-readable digital object to a state that is also machine-readable using the RDF data model, structured data can be created that is semantically interoperable, thus enabling the lexicographic resource to access, and be accessed by, other semantically interoperable resources.

Taking cognisance of ethical considerations in research, it was important to consider the copyright of EXDN. The copyright of literary works published in South Africa is regulated by the Copyright Act, No. 98 of 1978 ('the Copyright Act'). Section 3(1)(a) of the Copyright Act states that the author qualifies for copyright protection if the author "is a South African citizen or is domiciled or resident in the Republic"; however, Section 3(2) asserts that the copyright protection conferred on an author expires fifty years from the end of the year in which the person dies, with ownership of the literary work transferring to the public domain (*Copyright Act, No. 98 of 1978, as amended*, 2017:s3). As EXDN was published in South Africa and the author, Neil MacVicar, was domiciled in

South Africa for forty-seven years until his death in 1949 (Shepherd, 1952:214), EXDN is governed by the Copyright Act. More than sixty-five years have elapsed since MacVicar's death, so it is understood that copyright to EXDN is no longer held by the author and is consequently in the public domain, free from legal and copyright restrictions (Mitchell, 2013:12).

1.2.2 The structure of the dictionary

EXDN can be described according to its frame structure and its macrostructural and microstructural aspects. The frame structure of a dictionary typically comprises front matter texts, back matter texts, and a central list (Gouws & Prinsloo, 2005:57). In the case of EXDN, it comprises front matter texts and a central list only. The front matter texts consist of a title page, preface, abbreviations and shortened terms, weights, and measures. The central list represents the full list of the Roman alphabet, and it consists of a series of article stretches, with each letter of the alphabet serving as a guiding element.

The macrostructure of the dictionary consists of a lemmatised list of English terms with alphabetic ordering, with the nouns employing a singular and plural lemmatisation strategy. The microstructure of a dictionary refers to the structure of each lexical entry, typically represented by a lemma as the guiding element (Gouws & Prinsloo, 2005:119). For EXDN, each lexical entry (called an article) consists of a lemma and a definition or translation equivalent in isiXhosa. If a translation is equivalent, then there is a single target language item, represented by the word only (and not the stem), for which it is assumed there is full equivalence. If a definition, this may be interspersed occasionally with annotations, and it is assumed there is zero equivalence.

1.3 Conceptual framework

According to Miles and Huberman, "a conceptual framework explains, either graphically or in narrative form, the main things to be studied – the key factors, constructs or variables – and the presumed relationships among them" (1994:18). Maxwell talks of a conceptual framework as the context of a study (1996:25), and the

diagrammatic representation of this context "is a picture of the *territory* you want to study, not the study itself" (1996:37).

Represented in the form of a concept map, the context of the study is described in Figure 1-4.

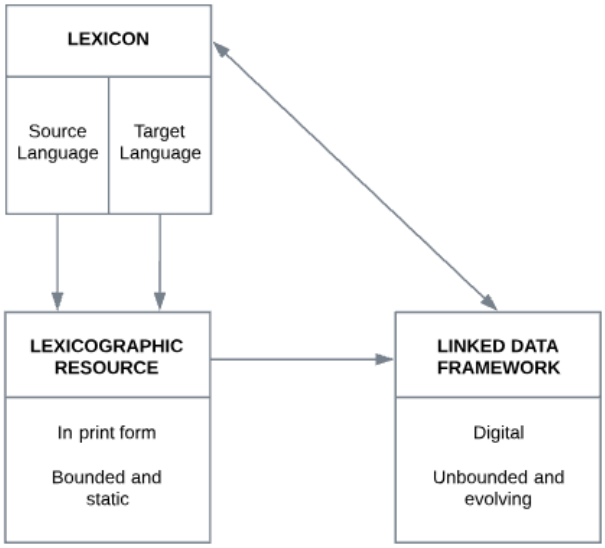


Figure 1-4: Concept map of the study's context

Drawing inspiration from the dendogram method (Miles & Huberman, 1994:251), each concept can be expanded further, as shown in Figure 1-5.

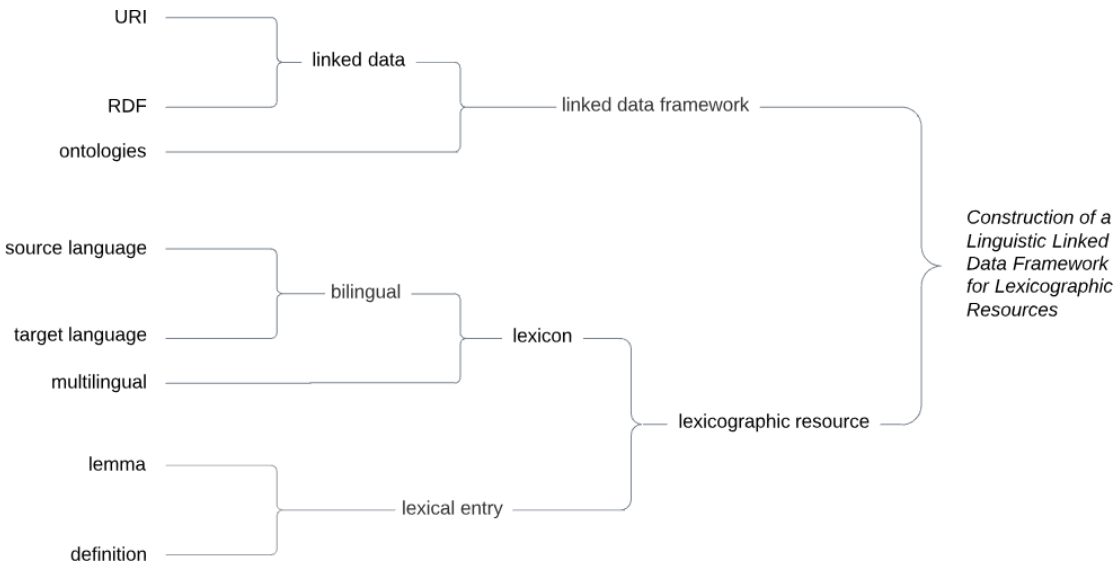


Figure 1-5: Expanding each concept

Below is a description of these key concepts and their related terms:

1.3.1 Linked data framework

An **ontology**, within the context of artificial intelligence and the web, is defined by Berners-Lee, Hendler and Lassila (2001:40) as "a document or file that formally defines the relations among terms." Gruber's definition, according to Stuart, is the most widely used definition of *ontology*: namely, "an explicit specification of a conceptualization" (Gruber, 1993:199; cited in Stuart, 2016:9); and the W3C, when introducing **vocabularies**, has iterated that there 'is no clear division between what is referred to as "vocabularies" and "ontologies"', with Hyvönen further explaining that although both refer to knowledge organisation systems (KOSs), *ontology* is the term "preferred for more complex and formal" KOSs (W3C, 2015; Hitzler, Krötzsch & Rudolph, 2010:46; Hyvönen, 2012:57).

An **ontological framework** may express data in the form of **RDF** triples and **URIs**, linking to ontological elements defined in ontologies and terms defined in vocabularies, and this concept is referred to as **Linked Data** (Coyle, 2012). For the purpose of this study, the ontological framework is referred to as a **linked data framework**, so that the term is not misinterpreted in a philosophical or more formal context. RDF has an **open-world assumption**, meaning that unless explicitly defined or declared, something cannot be assumed to be true (or false) as the information available may be insufficient to make this assumption (Van Hooland & Verborgh, 2014:61). RDF and Linked Data underpin the structure of the **Semantic Web**, defined by Berners-Lee as "a web of data"; and it is an extension of the existing web, enabling content that is **human-readable** to become **machine-readable** as well by making use of semantic markup (Berners-Lee, Hendler & Lassila, 2001:36; Berners-Lee, 2006; Coyle, 2012). **Linguistic Linked (Open) Data** (LL(O)D) refers to the publication of Linked Data for linguistics and natural language processing ("Linguistic Linked Open Data", 2018).

1.3.2 Lexicon

A **lexical item** is a unit of vocabulary, a word which can vary in form grammatically; however, it usually has a consistent meaning, for example, *abdomen* and *abdomens* of the

lexical item ABDOMEN (Crystal, 1997:221; Trask, 1993:158). If a lexical item is present in one language, but absent in another, this is referred to as a **lexical gap** (Crystal, 1997:221); however, if a term representing an identical or roughly identical concept exists in both languages, then this is an **equivalent** (Zgusta, 1971:312). The adoption of a word into a language is referred to as **lexicalisation** (Bussmann, 1996:279). A **lexicon** is defined by Crystal (1997:221) as "a complete inventory of the lexical items of a language [which] constitutes that language's dictionary, or lexicon." The mental lexicon, which will differ for each person, is the stored mental representation of lexical items of a language (Crystal, 1997:221). The term *lexicon* is synonymous with **vocabulary** (Crystal, 1997:221), although the latter term serves for everyday usage, and *lexicon* is reserved for more technical study. A distinction can be made between passive and active vocabulary, the former refers to lexical items which a person may understand but not use, whereas the latter is in use by the person (Crystal, 1997:411). Bussmann defines *vocabulary* as the "total set of all the words in a language at a particular point in time" (1996:514).

1.3.3 Lexicographic resource

Lexicography is defined by Bussmann as the "theory and practice of compiling dictionaries", and it "provides the principles necessary for documenting the vocabulary of a language" (1996:279). A **lexicographic resource** is a dictionary, and it lists the lexical items of a language, its lexicon (Crystal, 1997:221), although this list may be restricted depending on the domain and the target audience (Gouws & Prinsloo, 2005:47). In a dictionary, the lexical items listed are a set of **lexical entries** (Crystal, 1997:221). Each lexical entry is identified by a **lemma**, which is defined by Crystal as "an abstract representation, subsuming all the formal lexical variations which may apply" (2003:263). From this it can be inferred that the lemma "abdomen" is the written representation of the lexical item ABDOMEN, and the lexical entry is equivalent to a lexical item and its varying forms. **Lemmatisation** can be defined as the form of the lemma presented in a dictionary, and the approach or strategy taken may differ from one dictionary to another (Gouws & Prinsloo, 2015:67). When the lexical entry "'denotes' a particular object or state of affairs", this is **denotation as reference**; it is an **extensional reference**, and denotation is independent of context and situation (Bussmann, 1996:118).

As an example of **extension**, "tummy", "stomach", "venter" and "abdomen" all denote the stomach, although their intensional content differs (Bussmann, 1996:160). **Intension** is circumscribed by the **senses** of a lexical entry, where each sense has a different meaning, dependent on context and situation, and a sense includes the properties which define the lexical item (Crystal, 1997:198-199; Bussmann, 1996:160).

1.4 Research problem

EXDN is a bilingual dictionary for which English and Xhosa is the language pair; with English the source and isiXhosa the target language. IsiXhosa, an official language of South Africa, is an indigenous African language from the Nguni language group (Guthrie's S40) (Doke, 1954:91; "Subfamily: Nguni (S.40)", n.d.). Despite being spoken by a significant percentage of the population in South Africa, with 16.0% speaking it as a mother tongue[4] counted in the 2011 Census, isiXhosa enjoys minority status only (Statistics South Africa, 2012:24). When compared to English, there are limited linguistic resources available for isiXhosa; a scenario which applies to all indigenous African languages in South Africa (Herbert & Bailey, 2002:72; Pretorius, 2014:49-53).

According to Shepherd (1952:131), EXDN was written by MacVicar to serve as a teaching technique for nurses being trained at Victoria Hospital in Alice, in the Eastern Cape in South Africa. Despite more than eighty years having elapsed since the publication of the dictionary, its value as a lexicographic resource remains undiminished, particularly for L1 English-speaking nurses and doctors learning the African language as part of their hospital internship, whilst treating isiXhosa-speaking patients (Levin, 2014:290-291). Gouws and Prinsloo (2005:12) talk of users being "empowered by access to a dictionary", but in the instance of this dictionary, it is not just the users who are empowered, but indirectly, their patients as well. For patients, and by association, their families, being communicated with in their L1 not only aids

[4] Mother tongue is the L1; L2 is second-language acquisition – this may be a foreign language, or a local language that is not acquired as a mother tongue (Crystal, 2010:388).

understanding due to increased communicative success (Gouws, 1996), but also enables greater participation in any decision-making relating to the patient.

Official languages in South Africa in the Bantu language family (listed by their endonyms) are: isiNdebele, isiXhosa, isiZulu, Sesotho, Sesotho sa Leboa, Setswana, siSwati, Tshivenḓa, and Xitsonga. These languages are acknowledged to be under-resourced, due in part to the socio-economic constraints of the speakers and the limited language resources available, in the form of dictionaries, corpora and terminologies (Pretorius, 2014:50; ELRA, 2015). English is a lingua franca of South Africa, and although it is a colonial language with mother-tongue speakers in the minority (9.6%), it enjoys high status and is associated with political and economic power (Ngcobo, 2010:11; Statistics South Africa, 2012:24). Despite being spoken by the majority, the African languages listed above are minority languages and at risk of becoming endangered, through language shift and death (Pretorius, 2014:51; Ngcobo, 2010:16). Language is a symbol of social identity, with language and culture closely interlinked, and the loss of language leads to the loss of culture as well (Ngcobo, 2010:16).

Grover, van Huyssteen and Pretorius (2011:272, cited in Pretorius, 2014:53), in an audit on Human Language Technology (HLT) in South Africa in 2009, identified the "lack of language resources (LRs), limited availability of and access to existing LRs, (and) quality of LRs" as some of the common issues when developing LRs in under-resourced languages. Linked Data has an interoperable format and simple data model, and by making a lesser-known lexicographic resource such as EXDN available as Linked Data, it could be used as an aid in the development of future LRs, allowing for the "aggregation and integration of linguistic resources" (Gracia, 2017:17).

1.5 Research question

The primary objective of this study is to formulate a methodology for the construction of a framework for bilingual lexicographic resources, applying linguistic linked data principles. The bilingual lexicographic resource, *English-Xhosa Dictionary for Nurses*, was

used as a case study; the lexicon derived from the printed dictionary is the dataset, and English and isiXhosa is the language pair.

The research question which results from the primary objective is as follows:

How does one construct a framework for bilingual lexicographic resources, applying linguistic linked data principles? ***[Q0]***

To effectively address this question, sub-objectives have been identified, relating to (a) the extensibility of the framework, and (b) the representation of translation equivalents within the framework.

The linguistic linked data framework (LLDF) should not be constructed only for use cases identified in the dataset of the case study; instead, it should be able to extend to a multilingual resource should the data necessitate it. Therefore, the following sub-objective has been identified: allowing for extensibility from a bilingual to a multilingual resource.

This leads to the formulation of Q1 as:

How does one construct the LLDF so as to allow for extensibility from a bilingual to a multilingual resource? ***[Q1]***

Continuing with the theme of extensibility, the purpose of the LLDF is to define each lexical entry as "historically dynamic", instead of "ontologically static" (Veltman, 2006:6, cited in Rafferty, 2016:5). To achieve this for a framework which has instances in constant evolution, focus should be given to the provenance and linked data generation thereof.

This leads to the formulation of Q2 as follows:

How does one construct the LLDF to allow for change, tracking provenance of each change? ***[Q2]***

In the context in which this study is conducted, namely at a South African university, in South Africa, a previously colonised country, using a dataset with an indigenous African language as the target language of the language pair, it would be remiss to not consider the indigenous languages of South Africa when constructing the LLDF so as to ensure the varying forms of the languages can be represented.

Languages are said to be agglutinative when affixes (which are morphemes, the smallest unit of a language) are added to a word stem to create a word or phrase; the orthography (spelling system) of an agglutinating language can then be described as conjunctive or disjunctive (Crystal, 2010:303; Taljard & Bosch, 2006:428-429). Conjunctiveness refers to the affixes being bound together when written, disjunctiveness refers to the affixes being separated by whitespace (Taljard & Bosch, 2006:433). The official languages of South Africa in the Southern Bantu zone defined by Guthrie (cited in Herbert & Bailey, 2002:60-61) are agglutinative; however, there is a conjunctive or disjunctive orthography for each (Louw, 1984:231; Gouws & Prinsloo, 2012:78-79). As an example, isiXhosa has a conjunctive orthography, Sesotho sa Leboa is disjunctive (Gouws & Prinsloo, 2012:78-79; Taljard & Bosch, 2006:428-429).

A lemma is the address of a lexical entry which a person will use to retrieve information, and a word stem, word, or a multiword expression can all be lemmas in the same central list (Gouws & Prinsloo, 2005:64-67). The lexicographic tradition for the lemmatisation of nouns and verbs, namely word versus stem, will vary depending on the conjunctiveness or disjunctiveness of the language concerned (Gouws & Prinsloo, 2005:68). Due to this variation, the lemmatisation strategy for agglutinating languages should be considered when constructing the LLDF (Gouws & Prinsloo, 2008:75-84).

This leads to the formulation of Q3:

How does one construct the LLDF for translation equivalents, which may have a differing lemmatisation approach for nouns and verbs? ***[Q3]***

The Khoisan languages, although not accorded official language status (The Department of Justice and Constitutional Development, 2017), fall under the mandate of the Pan

South African Language Board (PanSALB), a Board brought into effect by the Constitution of South Africa (Act 106 of 1996), to "promote, and create conditions for the development and use" of the official languages, Khoisan languages, and South African Sign Language (SASL) (PanSALB, 2015). When constructing the LLDF, due consideration should thus be given to languages with click consonants which have additional letters in the alphabet, complementing the 26 letters of the Roman alphabet.

This leads to the formulation of Q4:

How does one construct the LLDF for lexical entries which may have letters not typically represented by the Roman alphabet? ***[Q4]***

As Q4 has cognate issues with Q1-Q3, it is included here but is not core to the investigation of the study. It will thus be dealt with superficially in Chapter 5 as an area that requires further study and research, when building upon the LLDF.

A proposal was put forth by Buitelaar et al. (2012:16) to consider the vision of the Semantic Web, i.e. this web of data, within the context of multilingualism, defining the Multilingual Semantic Web (MSW) as "the creation of a Semantic Web in which all languages have the same status, every user can perform searches in their own language, and information can be contrasted, compared and integrated across languages."

While this study will not be able to realise the vision of the MSW, it will contribute in a very small way by providing a framework in which lexicographic resources for under-resourced languages can be represented as Linked Data, with the lexical entries becoming machine-readable, thereby exposing the dataset to a far greater audience than would otherwise be possible. Figure 1-6 is a visualisation of the web of data, showing datasets published on the web in Linked Data format, referred to as the Linked Open Data Cloud (LOD cloud).

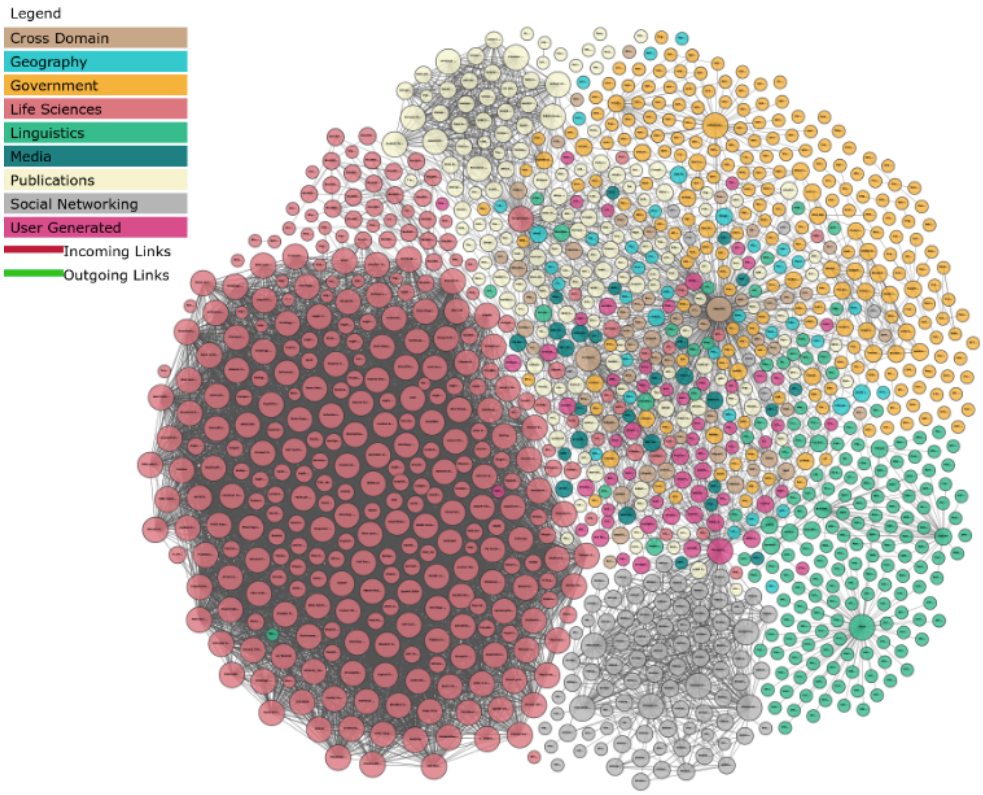


Figure 1-6: Linking Open Data cloud diagram (Abele et al., 2017)

A close-up view is given of the same diagram in Figure 1-7, showing the relationships between the sets of data, where each node is a dataset represented in RDF.

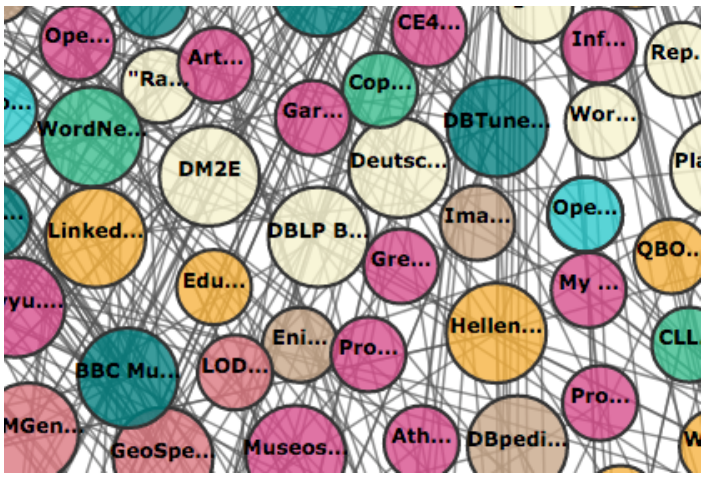


Figure 1-7: Close-up view of Figure 1-6

When a lexicon derived from a printed dictionary is converted to a dataset in Linked

Data format, it makes the data "shareable, extensible, and easily re-usable" ("Benefits", 2011), thereby creating the possibility for the lexicon, and the lexical entries contained therein, in the words of Berners-Lee (2009), to be "combined into something more interesting than the original pieces." However, the data are not just limited for use within other domain ontologies and linguistic resources; other practicable uses include HLT applications for the development of language resources, such as machine translation, multilingual comprehension assistants, and question answering (Grover, van Huyssteen & Pretorius, 2011:277).

1.6 Research approach

Figure 1-8 outlines the approach of this study, grouped into five sections, with each section corresponding to a chapter.

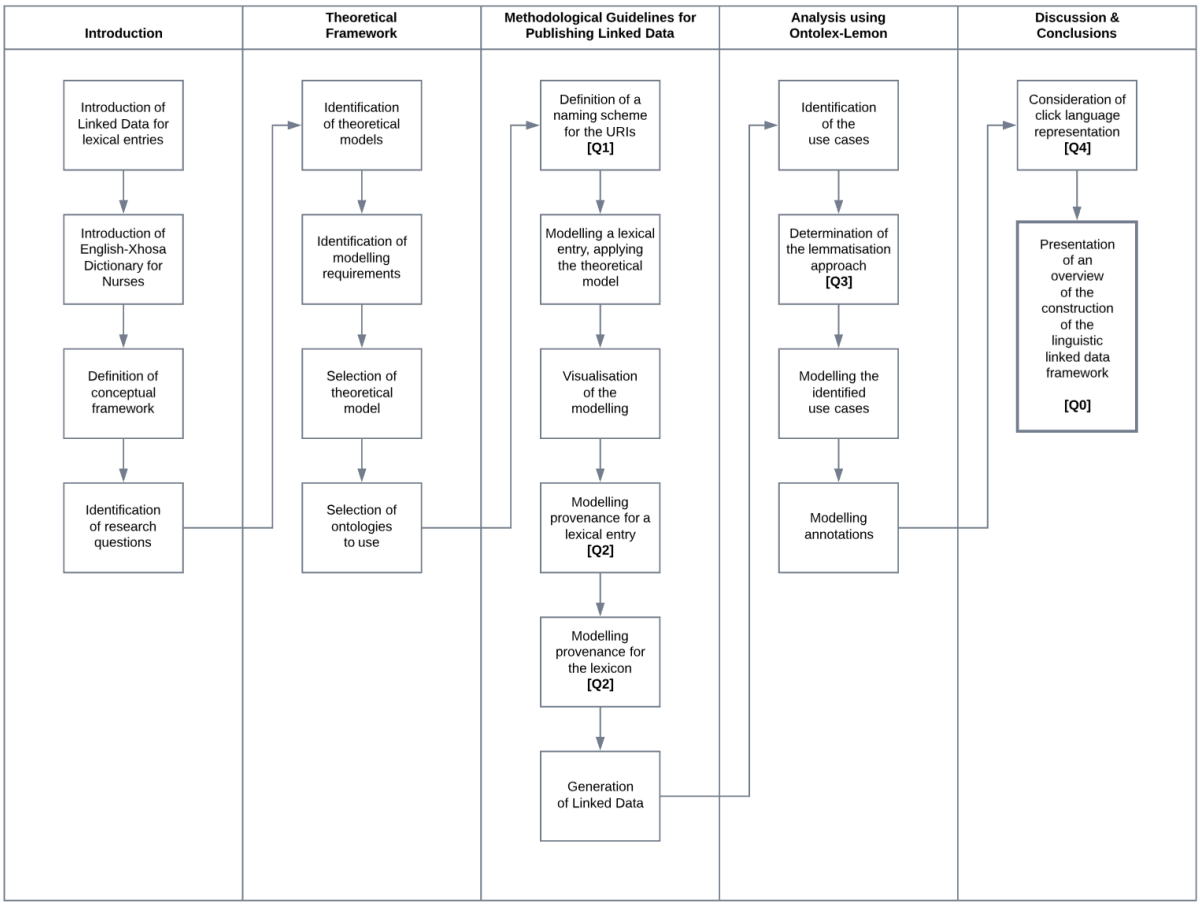


Figure 1-8: Overview of the research approach

Neither a qualitative nor quantitative research approach was adopted for this study, as both require analysis of the data collected, be it via observation or by measurement (Rasinger, 2013:50-52), which was deemed unsuitable for the topic. Instead, the case study as a research method was adopted: a bilingual lexicographic resource was used to test the methodological guidelines determined in Chapter 3 for the construction of an LLDF, using the model identified in Chapter 2 (Yin, 1994:27; Zainal, 2007). Although a single-case design was adopted, cognisance was taken of other use cases which would not be applicable to the single case study, notably regarding a resource with more than two vocabularies (Q1), and the orthographic diversity of other indigenous African languages in South Africa, as referred to by Zainal (2007) (Q3 and Q4).

The data derived from the lexicographic resource was imported into a MySQL database. A Dictionary Writing System (DWS) was custom-developed to enable the existing lexical entries to be maintained and new lexical entries to be added, and then for the lexicon to be published as RDF triples in Turtle and N-Triples format, using the methodological guidelines detailed in Chapter 3. The RDF dataset will be periodically uploaded to the following data centres: ZivaHub and Datahub.io. URIs using the project namespace are dereferenceable, with a valid RDF document being returned for each lexical entry ("How to contribute", n.d.). Providing a human-readable form of each lexical entry is outside the scope of this study, but the Turtle format should suffice in this regard (see Appendix B).

1.7 Limitations and delimitations

Limitations are defined by Simon (2011) as "potential weaknesses" in the researcher's study, which are beyond their control; delimitations are defined as "those characteristics that limit the scope and define the boundaries" of the researcher's study.

The following limitations have been identified:

- EXDN was published in 1935, and as a result, a number of lexical entries may refer to outdated or obsolete concepts . To mitigate this, when an entry is identified as

obsolete, the entry is excluded from the dereferenceable URIs and RDF dataset; and if outdated, the "outdated" property is modelled as such in RDF.

- Although the EXDN provides the isiXhosa equivalence of English terms, the work was primarily compiled by a person not of isiXhosa descent.
- The researcher has limited knowledge of isiXhosa and is unable to confirm the accuracy of every lexical entry and its translation equivalent or definition, or the quality thereof.

The following delimitations have been identified:

- Due to the size of the dataset and time constraints of the study, it is not possible to convert every lexical entry to RDF. This has been mitigated by abstracting the key characteristics of the lexical entries, and modelling the abstractions accordingly (Van Hooland & Verborgh, 2014:12).
- RDF has an open-world assumption, and as a result, the conversion of each lexical entry to RDF can never be regarded as fully complete; completeness is thus considered to be on a continuum (Van Hooland & Verborgh, 2014:61,161). To migitate this "incompleteness" for a framework which has instances in constant evolution, provenance and linked data generation is given a particular focus (Q2).
- As many structural issues exist in online lexicography for African languages, the researcher's focus is specifically on lemmatisation.

1.8 The report structure

This research report is divided into five chapters. The chapters are as follows:

Chapter 1: Introduction

The current chapter, this chapter has introduced the topic and background to the study. Various aspects, such as the lexicographic resource to be used as a case study, the research problem, the research questions, and the research approach have been covered in detail.

Chapter 2: Literature Review & Theoretical Framework

This chapter considers theory and the case study approach within the context of this study. In the absence of a formal theoretical framework, various models which can inform the study are reviewed.

Chapter 3: Methodological Guidelines for Publishing Linked Data

This chapter presents the methodological guidelines for the publication of Linked Data, and the methodology for the construction of the LLDF, using the selected model from Chapter 2. The research questions Q1 and Q2 are addressed in this chapter.

Chapter 4: Analysis using Ontolex-Lemon

This chapter describes the application of the methodological guidelines from Chapter 3, using EXDN, described in Chapter 1, as the case study. The research question Q3 is addressed in this chapter as well.

Chapter 5: Discussion & Conclusions

This chapter concludes the study, summarising the findings from Chapters 2, 3 and 4. The research question Q4 is addressed, with the research question Q0, which relates to the primary objective of the study, discussed in detail. The researcher ends the chapter with recommendations for future work.

Chapter 2 Literature Review & Theoretical Framework

2.1 Introduction

"Theory," Pierce declared, "is not dry abstraction but the body of concerns, methods and research problems a discipline develops over time" (1992:641-643, cited in Ocholla & Le Roux, 2011:5). Maxwell has defined theory as "a set of concepts and ideas and the proposed relationships among these, a structure that is intended to capture or model something about the world" (2013:48). He goes on to say that the "simplest form of theory consists of two concepts joined by a proposed relationship" (2013:49); this is not dissimilar to a subject and an object, joined by a predicate… an RDF triple, so to speak. Theory is the simplification of observed relations of a phenomenon (Maxwell, 2013:49; Anfara & Mertz, 2006:xvii), be it something large, for example, the extinction of the dinosaurs, or small - such as the representation as Linked Data of an alphabetised list of lexical entries in a bilingual dictionary published in the 20th century.

2.1.1 Case study approach

As outlined in Section 1.6, a case study approach was adopted. Eisenhardt defines the case study as "a research strategy which focuses on understanding the dynamics present within single settings" (1989:534). Yin (2014:16) has defined the case study approach as:

> an empirical inquiry that
>
> - investigates a contemporary phenomenon (the 'case') in depth and within its real-world context, especially when
> - the boundaries between phenomenon and context may not be clearly evident.

By using the case study approach, the representation of lexical entries from EXDN as Linked Data (the phenomenon) within a specific context, namely South Africa and its indigenous African languages, can be examined closely (Zainal, 2007:1-2). Yin has argued that the theoretical perspective should be identified at the outset as it impacts the research questions; for this study however, the identification of the theoretical

perspective prior to determining the research questions was not deemed necessary, although the identification of the context and the 'boundedness' of this context was. Various aims can be accomplished with the case study approach: it can be used to generate theory, to test theory, and to provide description (Eisenhardt, 1989:535). Woodside and Wilson have suggested that the quality of a case study report "often may be increased dramatically" if the study allows for both the testing of theory and to generate theory (2003:502).

According to Burns (1997:364, cited in Kumar, 2014:155), a "case study should focus on a bounded subject/unit that is either very representative or extremely atypical", and this bounded subject need not be a singular object such as an individual person or an enterprise (Stake, 2000:23). The dataset of this study is bounded, limited to the lexical entries contained within EXDN, the physical artefact. Although this is a single-case study, the single case consists of its own cases, namely the lexical entries, and within each lexical entry, one or more characteristics can be identified. Despite the study being bounded, the lexical entries are of a sufficiently large number that it was not possible to convert each entry to RDF. As mentioned in Chapter 1, the key characteristics were identified as abstractions of reality, and the abstractions modelled; the findings were then generalised to the entire case as per Van Hooland and Verborgh (2014:12) and Gomm, Hammersley and Foster (2000:103). The key characteristics were not selected by means of random sampling; instead, they were purposively selected as "suitable for illuminating and extending relationships and logic among constructs" – necessary when one is generating theory (Eisenhardt & Graebner, 2007:27). Although the findings are generalised to the entire case, they are not limited to the single case; instead they are generalised to the theory, and these generalisations can then be applied to different contexts (Yin, 2014:40-42; Tellis, 1997:103). Despite a single case design being adopted, the context of the original lexical entries, namely English as the source language and isiXhosa as the target language, was expanded to allow for the representation of lexical entries in other indigenous African languages of South Africa. While Afrikaans is 'indigenous' in the sense that it is not spoken outside of southern Africa, it belongs to the West Germanic language family, and as such, English is considered sufficiently representative (Mesthrie, 2002:5-6; "Subfamily: West Germanic", n.d.); indigenous

African languages outside of South Africa were also not considered as the researcher felt the research questions could be adequately applied to these languages. This bounding of the context was done to avoid "theoretical saturation": the point at which the addition of new cases, or for this study, an expanded context, stops contributing to the building of theory (Eisenhardt, 1989:545).

2.1.2 Literature review and theoretical framework

Just as the selection of cases is an essential component of theory building, so is the literature review (Eisenhardt, 1989:536,544). The review of the extant literature throughout the research process is invaluable: in the early stage it helps to scope the research problem, providing a theoretical background to the study, and in the latter stages it is integral to achieving the aims of the case study approach, namely to test and generate theory, and to provide description (Kumar, 2014:48; Eisenhardt, 1989:535-536).

Another essential component to theory building is the theoretical framework. Ocholla and Le Roux define the theoretical framework as "the agenda, outline, and theoretical construct of a research approach" which "normally precedes the literature review", and it is "the structure that holds and supports the theory of the research work" (2011:1). Merriam (1998, cited in Anfara & Mertz, 2006:xxiii) calls the theoretical framework "the structure, the scaffolding, the frame of your study." Within Library and Information Science (LIS), a theoretical framework unique to the subject is not available, with researchers instead using theories from other disciplines (Ocholla & Le Roux, 2011:1-5). Pierce is quite critical of this, referring to researchers "seeking favour by imitating practices of disciplines considered superior to its own" (1992:641-643, cited in Ocholla & Le Roux, 2011:5). Considering the interdisciplinary nature of LIS, with its focus on human knowledge, and its increasing digitalisation and interconnectedness (Simons & Richardson, 2013:12), using multi-disciplinary research frameworks and models from neighbouring disciplines is not wholly unreasonable (Ocholla & Le Roux, 2011:5-10).

When one considers the physical artefact, EXDN – it is closed and bounded; in contrast, the web is open and unbounded (Di Maio, 2015:3). Data published as Linked Data, although bounded by both its namespace and its licensing, by using HTTP URIs, operates on top of this open web (Bizer, Heath & Berners-Lee, n.d.). Di Maio (2015:10)

talks of knowledge unification (although the researcher prefers the term 'interconnectedness'), referring to Linked Data as "the nearest publicly available artefact ever to make knowledge unification a *de facto* reality." At an abstract level, General Systems Theory, which deals with the general properties of systems, can elaborate on generalised models of systems, where these systems can "serve to describe nature and our existence" (Skyttner, 1996:24) but according to Anfara and Mertz (2006:194-195), "no theoretical framework adequately describes or explains any phenomena", where they describe the power of a theoretical framework as "too reductionistic" and "too deterministic." Ocholla and Le Roux have stated that the literature review is a theoretical framework in itself (2011:7), and in the absence of a theoretical framework which can sufficiently inform this study beyond the abstract level, a model, as well as the literature review, have sufficed. When a literature review is conducted, Kumar recommends writing up the findings of the literature, organised according to the main themes which emerged during the literature search (2014:50). However, for this study, a systematic approach has been taken; the structure of the literature review follows that of Onwuegbuzie and Frels (2016:31):

- A review is undertaken of the available models (Section 2.2).
- Once a model has been selected, a review of the methodological guidelines for publishing Linked Data is undertaken (Section 3.1).
- Once the methodological guidelines have been reviewed, each step identified in the guidelines serves as a theme, with a review conducted for each (Sections 3.2 onwards).

2.2 Models for Linguistic Linked Data

As previously defined, an ontology is "an explicit specification of a conceptualization" (Gruber, 1993:199, cited in Stuart, 2016:9), and at a coarse-grained level, a lexicon can share this definition, with ontologies and lexical resources sharing enough similarities that they are sometimes "used interchangeably or combined into merged resources" (Prévot et al., 2010:4-5). However, as argued by Hirst (2004, cited in Prévot et al.,

2010:5), lexicons are "not really ontologies" as formal ontologies are supposed to be grounded in unambiguity with synonym terms grouped under the same concept, whereas for lexicons, synonymy and near-synonymy are important relations which do not necessarily share the same concepts (Prévot et al., 2010:5), so although lexical resources and ontologies are "objects of the same nature", they differ with regards to "conceptualization, specification and scope" (Prévot et al., 2010:9). Extending the lexical context of ontologies, or extending the ontological representation of semantics of entries in lexicons, has led to the proposal of models for representing this *ontology-lexicon interface* (McCrae & Unger, 2014:15; Prévot et al., 2010:9-11), with the ontology forming a "shared conceptualisation" and the lexicon describing the "lexical encoding of that conceptualisation in words" (McCrae & Unger, 2014:26).

In 2017, the 2nd Summer Datathon on Linguistic Linked Open Data[5] was held in Spain ("2nd summer datathon …", n.d.). As a datathon series held biennially, focussing on the field of language resources and "unique in its topic worldwide", its purpose was to enable participants to migrate their linguistic data from an existing data source and publish it as Linked Data on the web ("2nd summer datathon …", n.d.). At this datathon, Ontolex-Lemon[6] was presented as the principal model for representing Linguistic Linked Data (LLD).

Ontolex-Lemon was initially published by the Monnet project in 2010 as *lemon* – the Lexicon Model for Ontologies[7], and in May 2016, *lemon* was published as a W3C vocabulary[8], now referred to as Ontolex-Lemon, and this model remains under active development by the W3C Ontology-Lexica Community Group ("lemon – the lexicon …", n.d.; Cimiano, McCrae & Buitelaar, 2016). The Ontolex-Lemon model (and thus the *lemon* model) represents lexicons and machine-readable dictionaries "relative to ontologies by a principle called *semantics by reference*", defined as a case in which "the meaning of a word is given by reference to an ontology, resulting in a clean separation between the lexical and semantic layer" (McCrae & Unger, 2014:16).

[5] http://datathon2017.retele.linkeddata.es/
[6] https://www.w3.org/community/ontolex/wiki/Final_Model_Specification
[7] http://lemon-model.net/
[8] https://www.w3.org/2016/05/ontolex/

lemon was influenced by Lexical Markup Framework[9] (LMF), as well as the LexInfo[10] and Linguistic Information Repository[11] (LIR) models (Cimiano, McCrae & Buitelaar, 2016; McCrae, 2012; McCrae & Unger, 2014:27).

LMF is ISO standard 24613:2008, designed between 2003 and 2008, and intended to provide a standardised framework for natural language processing (NLP) and machine-readable dictionaries (Francopoulo & George, 2013:19). Although able to represent linguistic information, it is unable to represent lexicons to ontologies (McCrae, Spohr & Cimiano, 2011:3, McCrae, 2012), and despite describing itself as interoperable, LMF has been criticised for its inability to establish interoperability between different lexicons and its vagueness for use when applied to different contexts (McCrae & Unger, 2014:27; Faab, Bosch & Gouws, 2014:96). However, the *lemon* model (on which Ontolex-Lemon was founded) was inspired by LMF, with *lemon* adopting its core ontology, importing classes and entities from LMF but adding vocabulary in order to describe the ontology-lexicon interface (Cimiano, McCrae & Buitelaar, 2016).

LexInfo proposed a model that unified LMF with the OWL ontology model, by conceptually building on three components: the LingInfo and LexOnto models and LMF (McCrae & Unger, 2014:27; Cimiano et al., 2010:30). The LingInfo[12] and LexOnto models were complementary, with the former providing a mechanism "for modelling label-internal linguistic structure", such as inflection, interpreted as terms, and the latter enabling "the representation of label-external linguistic structure", with mappings to ontological structures (Cimiano et al., 2010:30). By combining aspects of both models, LexInfo enabled linguistic information (such as part of speech (POS) and inflection) to be associated with ontology elements (such as concepts, relations, instances) in a way that was reusable across systems (Cimiano et al., 2010:29-30). While RDF and RDF Schema (RDFS), Web Ontology Language (OWL) and Simple Knowledge Organization System (SKOS) can be used to associate labels with ontology elements, these do not describe the linguistic information thereof, although SKOS does allow for further

[9] http://www.lexicalmarkupframework.org/
[10] http://www.lexinfo.net/
[11] http://mayor2.dia.fi.upm.es/oeg-upm/index.php/en/technologies/63-lir/
[12] http://olp.dfki.de/LingInfo/

typology, such as identifying the label as "preferred" (Cimiano et al., 2010:29-30; Vila-Suero et al., 2014:110).

LIR made use of an OWL meta-ontology which can be associated with any element of an OWL ontology, and it focuses on the variations of terms (such as acronyms and transliterations), explicitly defining translation relations between these term variants (Montiel-Ponsoda et al., 2011:106; Espinoza, Gómez-Pérez & Montiel-Ponsoda, 2009:822).

Figure 2-1 shows a timeline of the models under discussion (as at December 2017).

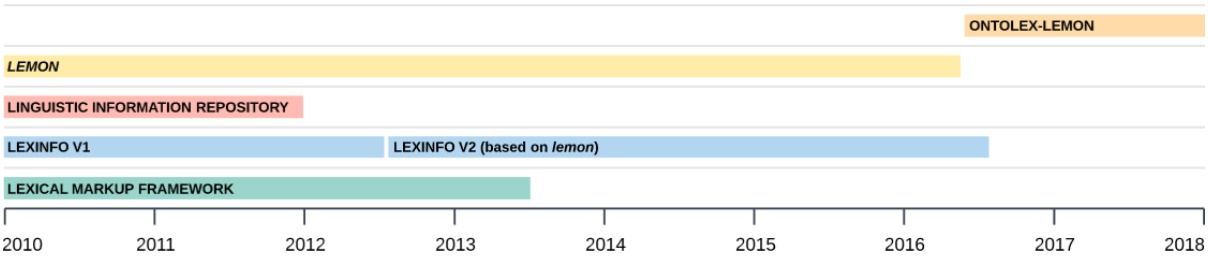


Figure 2-1: Timeline of the models (in active development)

In the literature reviewed from 2012 to 2017, LMF, LexInfo, *lemon*, and Ontolex-Lemon have served as models for lexical resources.

2.3 Modelling requirements

The modelling requirements are broadly defined in six sections – with the requirements aligning closely with the data of EXDN (Cimiano et al., 2010:29-33):

1. Interoperability

Does the model support interoperability, building on existing standards for RDF?

2. Separation and independence

Does the model allow for separation between the lexical and the ontological layer, where linguistic information is modelled in a separate ontology?

For example: An external resource can link to the same semantic layer (expressed using an ontology), but use a different lexicon, thereby allowing for extensibility and reuse.

3. Linguistic information

Does the model allow for structured linguistic information to be captured?

For example: Identifying an ontology class with a word expressed in natural language, and then identifying the plural of that word, or identifying the POS (Cimiano et al., 2010:31).

4. Morphological decomposition

Does the model allow for words to be decomposed into their smaller parts (morphemes)?

For example: Separating the word stem from its affixes for isiXhosa lexical entries.

5. Multilinguality

Does the model support multilingualism and the association of translation relations, beyond language tagging?

For example: Explicitly declaring an isiXhosa lexical entry to be the translation equivalent of an English lexical entry.

6. Ontological representation

Does the model allow for ontology entities to serve as a representation of meaning?

For example: Selecting a DBpedia resource as the ontology entity which serves as the denotation of the lexical entry.

7. Linked Data principles

Does the model adhere to basic Linked Data principles?

Table 2-1 shows a comparison of each model according to the modelling requirements. SKOS is included for informational purposes.

	1 Inter-operability	2 Separation & Independence	3 Linguistic Information	4 Morphological Decomposition	5 Multi-lingualism	6 Ontological Representa.	7 Linked Data Principles
SKOS	Yes	No	No	No	No	No *	Yes
LMF	No	No	Yes	Yes	Yes	No	No
LexInfo	Yes	Yes	Yes	Yes	No	Yes	Yes
LIR	Yes	Yes	Yes	No	Yes	Yes	Yes
Ontolex-Lemon	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Table 2-1: The models' features according to the modelling requirements

(The table has been derived from Cimiano et al. (2010:33), updated to include Ontolex-Lemon, and adapted to the modelling requirements listed above. *Cimiano et al. indicates this as "Not applicable".)

2.4 The selected model

As shown in Figure 2-1 and Table 2-1, as well as meeting all the modelling requirements, Ontolex-Lemon is the only model under active development. Ontolex-Lemon has thus been selected as the model to use for this study.

Ontolex-Lemon builds on the *ontology-lexicon interface* paradigm, employing the use of the *semantics by reference* principle, both described in Section 2.2, with the separation of the ontological and lexical layers; the advantage of this is that by changing its lexicon, an ontology can change from one language to another language (McCrae et al., 2017:587). It consists of the core module, shown in Figure 2-2, and the additional modules: Syntax and Semantics (*synsem*), Decomposition (*decomp*), Variation and Translation (*vartrans*), and Metadata (*lime*), can be used as required (Cimiano, McCrae & Buitelaar, 2016).

The primary element in the core module is the *Lexical Entry*, which can represent a single word, multiword expression, or affix (McCrae et al., 2017:589). The meaning of a lexical entry is given by reference to an ontology entity, and lexical senses can be defined for a lexical entry as well (McCrae et al., 2017:589). Modelling the requirements [1], [2], [3],

[4], [6] and [7] can be achieved with the core module; [5] can be modelled using the *vartrans* module.

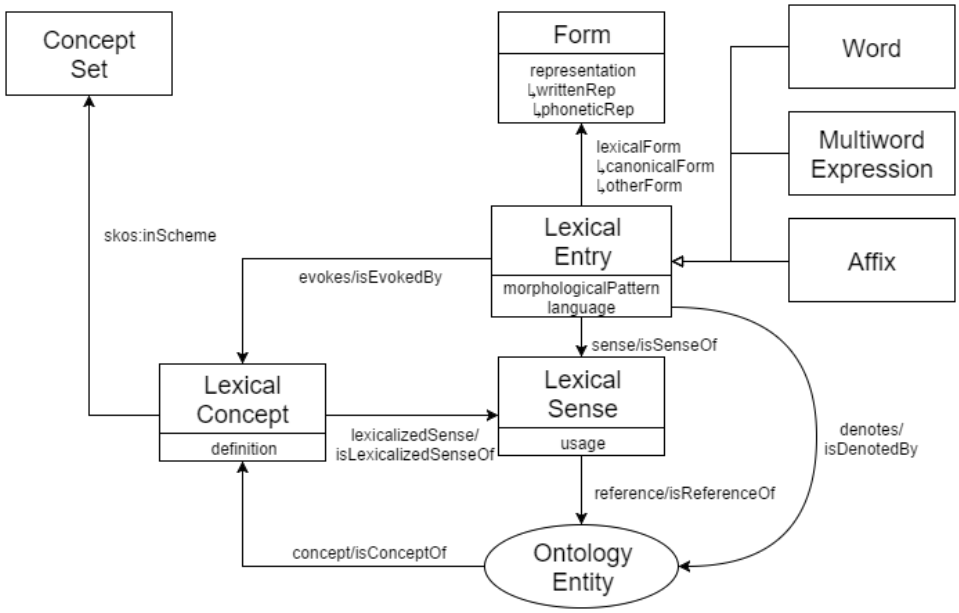


Figure 2-2: Ontolex-Lemon core module (Cimiano, McCrae & Buitelaar, 2016)

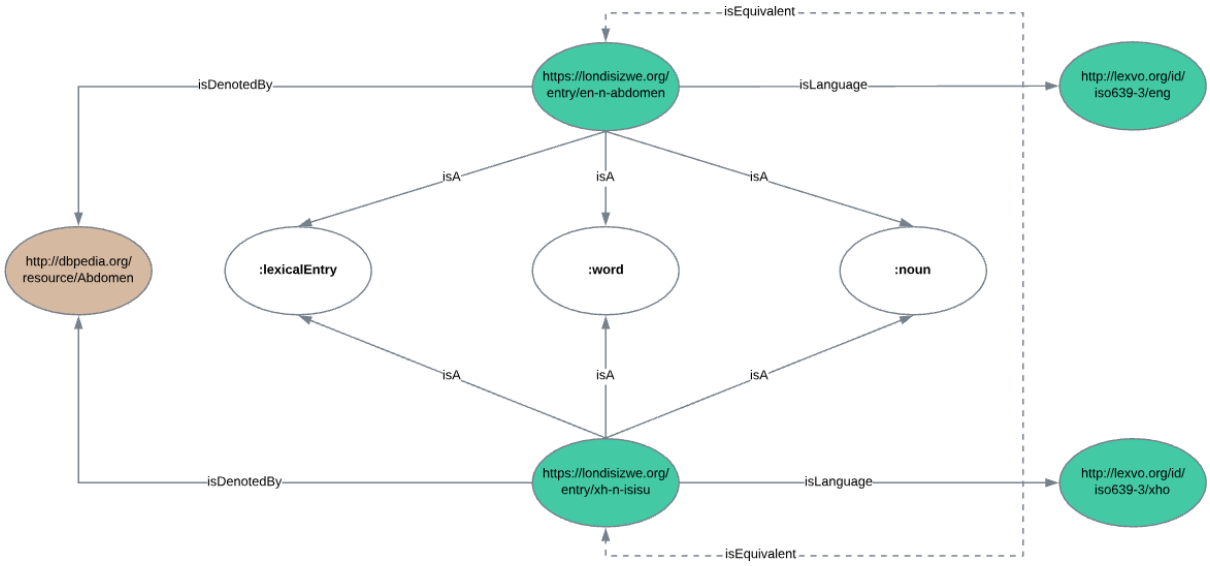


Figure 2-3: Illustrating the semantics by reference principle

To illustrate the *semantics by reference* principle, for the two lexical entries *abdomen* and

isisu, described in Section 1.2, Figure 2-3 shows that if the same ontology entity is identified to denote the meaning of both lexical entries, then their equivalence can be inferred.

2.5 Similar studies

Both the *lemon* and Ontolex-Lemon models have enjoyed widespread use, and examples of language resources converted to RDF include the Apertium Bilingual Dictionaries[13], the German monolingual dictionary in K Dictionary's Series, the Pattern Dictionary of English Verbs[14], and the classical Al-Qamus dictionary (2017:590-591); BabelNet, when converted to Linked Data, also used the *lemon* model (Ehrmann et al., 2014:402). Three dictionaries are considered here in more detail, namely the *'Al-Qāmūs Al-Muhit*, the *Apertium Bilingual Dictionaries*, and *Dictionnaire étymologique de l'ancien francais*.

2.5.1 ''Al-Qāmūs Al-Muhit'

''Al-Qāmūs Al-Muhit' (AQAM) is a print dictionary of Classical Arabic which was digitised and encoded in RDF using *lemon*, ensuring the digitised resource conformed to the printed form of AQAM (Khalfi, Nahli & Zarghili, 2016:325). Lexical entries are ordered "alphabetically based on the last radical consonant", with each consonant serving as the article stretch, referred to as "chapters" by the authors, and with each chapter further divided into sub-chapters, each according to the first consonant (Khalfi, Nahli & Zarghili, 2016:325). Senses in a lexical entry are linked to external resources (where available) and these resources include Arabic WordNet, Princeton WordNet (PWN), Suggested Upper Merged Ontology (SUMO) and an Arabic-English bilingual dictionary (Khalfi, Nahli & Zarghili, 2016:328). The result is a digital resource of AQAM which "will help Arabic language studies to gain on several fronts (lexicography, semantics, philology …)" (Khalfi, Nahli & Zarghili, 2016:329).

[13] http://linguistic.linkeddata.es/apertium/
[14] http://pdev.org.uk/PDEVLEMON.html

2.5.2 The Apertium Bilingual Dictionaries

The Apertium Bilingual Dictionaries (ABD) is a list of lexicons consisting of twenty-two linguistic datasets representing Spanish (Castillian), French, Italian and Portuguese, as well as other under-resourced languages, examples of which include Occitan, Asturian and Aragonese (Gracia et al., 2018:1). The linguistic datasets consist of lexicons derived from WordNets, such as Galician EuroWordNet-Lemon lexicon, or lexicons with language pairs, such as Catalan and Spanish, derived from the Spanish-Catalan LMF Apertium Bilingual Dictionary (Gracia et al., 2018:1; "Lexica", n.d.). Each dictionary was converted from a single LMF file into three RDF resources: the source and target lexicons, and the translation set between both lexicons (Gracia et al., 2018:4). By publishing the dictionaries as Linked Data using the *lemon* model, the result is the emergence of a multilingual dictionary, for which both direct translations and indirect translations (by means of an intermediary language) can be obtained from "a large unified graph of linked lexical entries, senses and translations" (Gracia et al., 2018:7).

2.5.3 Dictionnaire étymologique de l'ancien francais

The *Dictionnaire étymologique de l'ancien francais* (DEAF) is an ongoing dictionary project whose purpose is to document and study the Old French language (from the first-published resource in 842 CE until 1350 CE), and published resources of DEAF include printed books and an electronic dictionary (Tittel & Chiarcos, 2018:1). As a proof of concept to test the conversion of the dictionary data from XML to RDF, an exemplar dictionary article from DEAF was transformed into RDF, using Ontolex-Lemon. Senses in the exemplar were linked to external resources, and the original information from the accompanying DEAF article were also included (Tittel & Chiarcos, 2018:3). The result was the publication of "a novel set of philological lexical data in compliance with Linked Data principles", which could then contribute as a dataset to the LOD cloud shown in Figure 1-6; this data was then used to enrich a medical treatise written in medieval French, with the inclusion of references to DEAF's electronic dictionary (Tittel & Chiarcos, 2018:3).

The conversion of these three dictionaries to RDF using Ontolex-Lemon (or its predecessor, *lemon*) validates by precedence the conversion of EXDN to Linked Data

using Ontolex-Lemon. Furthermore, the conversion of EXDN from a printed dictionary to Linked Data within the context of a multilingual South Africa differentiates the project sufficiently from the studies discussed here.

2.6 Ontologies and vocabularies

Ontologies and vocabularies provide a structured framework to represent knowledge. This can be pertinent to a single domain, or it can be a general representation common to all domains (Arp, Smith & Spear, 2015:38). The models for representing ontologies and vocabularies in the Semantic Web space are RDFS, SKOS, and OWL (Hyvönen, 2012:63).

2.6.1 RDFS

RDFS was developed to augment the expressivity of RDF by introducing object-oriented modelling, with a domain described in terms of classes, with instances belonging to those classes, and with properties describing both the classes and the instances (Hyvönen, 2012:63; Arp, Smith & Spear, 2015:154). Notably, it introduced *domain* and *range* as property constraints, and this can be explained by means of an example:

Subject	*Predicate*	*Object*
"Neil MacVicar"	`dct:creator`	"English-Xhosa Dictionary for Nurses"

where the property `dct:creator` has the domain `Person` and the range `Document`, and domain applies to the Subject and range to the Object, respectively (Hyvönen, 2012:63).

2.6.2 OWL

OWL extends RDFS, by adding more vocabulary with a formal semantics to describe properties and classes, with ontological concepts able to be precisely defined (Hyvönen, 2012:64; W3C, 2004b; Van Hooland & Verborgh, 2014:126). It is a computational logic-based language that enables ontology-based reasoning and data validation, with three versions that provide increasing expressive power: OWL Lite, OWL DL, and OWL Full (Hyvönen, 2012:64). In the context of Linked Data, vocabularies can be expressed

sufficiently in RDFS, using primitives from OWL, such as `owl:sameAs` (Heath & Bizer, 2011:57). Cardinality, Boolean operators and equivalence can be expressed, as well as property characteristics: inverseOf, transitive, functional (cardinality is maximum of 1) and inverse-functional (Hyvönen, 2012:64; W3C, 2004b).

2.6.3 SKOS

This is a lightweight RDFS and OWL-compatible ontology format for the representation of vocabularies, where instances of concepts or collections of concepts can be connected as a semantic network (Hyvönen, 2012:63-64). The basis of SKOS is concepts, not terms; concepts can be defined using `skos:Concept` and hierarchical relations can be represented using `skos:broader` and `skos:narrower`, with properties for defining equivalence also available (Hyvönen, 2012:63-64; Van Hooland & Verborgh, 2014:130).

2.6.4 Selected ontologies and vocabularies

The ontologies and vocabularies identified for use in EXDN are described in Table 2-2.

2.6.5 Created vocabularies

The use cases of the case study could not be modelled sufficiently using external ontologies and vocabularies, so two vocabularies were created to assist with this, described in Table 2-3.

2.6.6 Defining the namespaces

In the Turtle serialisation, a prefix label can be declared for each of the repeating URIs (W3C, 2014d), making the URIs more readable. The prefix declarations for each of the ontologies and vocabularies discussed in this study are listed in Code Example 2-1. For the examples in the chapters that follow, it is assumed that the prefixes have been declared.

NAME SPACE		
dbr	DBpedia	A cross-domain ontology used to identify resources (DBpedia, 2018). For example: `dbr:Abdomen`
dct	Dublin Core Metadata Initiative	A set of vocabulary terms used to describe the properties of resources (Dublin Core Metadata Initiative, 2012). For example: `dct:creator`
foaf	FOAF	A dictionary of terms used to describe properties and identify resources (Brickley & Miller, 2014). For example: `foaf:Document`
lcnaf	Library of Congress Name Authority File	A controlled vocabulary to identify persons, organisations, etc. ("Library of Congress Names", n.d.). For example: `lcnaf:n87888720`
lcsh	Library of Congress Subject Headings	A controlled vocabulary to categorise resources ("Library of Congress Subject …", n.d.). For example: `lcsh:sh85000091`
lexinfo	LexInfo Vocabulary	A vocabulary which builds on the *lemon* model, and represents lexical information (Wunner, 2012). For example: `lexinfo:Noun`
mesh	Medical Subject Headings	A controlled vocabulary used to identify and categorise resources in the medical domain (U.S. National Library of Medicine, 2018). For example: `mesh:D000005`
mmoon	MMoOn	A multilingual morpheme ontology for expressing linguistic concepts and relations ("The Multilingual Morpheme …", 2018). For example: `mmoon:Stem`
prov	PROV Ontology	An OWL ontology used to represent provenance information (W3C, 2013b). For example: `prov:generatedAtTime`
pwn	Princeton WordNet 3.1	An RDF interface for Princeton WordNet ("WordNet RDF", n.d.). For example: `pwn:00836693-n`
void	VoID Vocabulary	A vocabulary for expressing metadata about datasets (Alexander et al., 2011). For example: `void:Dataset`

Table 2-2: Selected ontologies and vocabularies used

NAME SPACE		
lonvoc	Londisizwe Noun Class Vocabulary	A vocabulary formalised in OWL, describing the noun classes of African languages (starting with isiXhosa). See Appendix E. For example: `lonvoc:IsiXhosaNC7`
loncon	Londisizwe Concepts for Senses	A vocabulary for describing concepts to which senses are associated, intended to be a standalone inventory. For example: `loncon:000000001`

Table 2-3: Created vocabularies

prefix_declarations.ttl — ~/Documents/Personal/UCT/LIS5031W/Files/final

prefix_declarations.ttl

```
1
2  @prefix ontolex:  <http://www.w3.org/ns/lemon/ontolex#> .
3  @prefix lime:     <http://www.w3.org/ns/lemon/lime#> .
4  @prefix vartrans: <http://www.w3.org/ns/lemon/vartrans#> .
5
6  @prefix rdf:      <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
7  @prefix rdfs:     <http://www.w3.org/2000/01/rdf-schema#> .
8  @prefix skos:     <http://www.w3.org/2004/02/skos#> .
9  @prefix owl:      <http://www.w3.org/2002/07/owl#> .
10 @prefix xsd:      <http://www.w3.org/2001/XMLSchema#> .
11
12 @prefix dct:      <http://purl.org/dc/terms/> .
13 @prefix foaf:     <http://xmlns.com/foaf/0.1/> .
14 @prefix lexinfo:  <http://www.lexinfo.net/ontology/2.0/lexinfo#> .
15 @prefix mmoon:    <http://mmoon.org/core/> .
16 @prefix prov:     <http://www.w3.org/ns/prov#> .
17 @prefix void:     <http://rdfs.org/ns/void#> .
18
19 @prefix dbr:      <http://dbpedia.org/resource/> .
20 @prefix lcnaf:    <http://id.loc.gov/authorities/names/> .
21 @prefix lcsh:     <http://id.loc.gov/authorities/subjects/> .
22 @prefix mesh:     <http://id.nlm.nih.gov/mesh/> .
23 @prefix pwn:      <http://wordnet-rdf.princeton.edu/rdf/id/> .
24
25 @prefix lonvoc:   <https://ontology.londisizwe.org/nounclass#> .
26 @prefix loncon:   <https://concept.londisizwe.org/> .
27
```

prefix_declarations.ttl 2:1 (1, 58) LF UTF-8 Turtle 0 files

Code Example 2-1: Prefix declarations

Where:

- Line 2: `ontolex` is the core module of Ontolex-Lemon
- Lines 3-4: `lime` and `vartrans` are the additional modules of Ontolex-Lemon

2.7 Summary

If the underlying data structure of a dataset changes, a principled model such as Ontolex-Lemon instead of an RDF schema designed specifically for the dataset, allows the RDF schema to remain unchanged, thus allowing for extensibility to other datasets (McCrae, Montiel-Ponsoda & Cimiano, 2012:33). However, the *semantics by reference* principle is not without its drawbacks – if one considers the concepts *breath* (noun) and *breathing* (verb), both are denoted in DBpedia by `dbr:Breathing`, yet this is an inaccurate representation for a single breath. Another issue, identified by Hirst (2014:5), is the lack of equivalence between words in multilingual contexts. Both of these issues can be considered limitations of the model, which will be considered in the following chapters.

This chapter considered theory within the context of LIS, and the boundaries of this study. The different models for representing lexical data were reviewed, and the modelling requirements identified, culminating in the selection of the Ontolex-Lemon model for representing LLD. The following chapter details the methodological guidelines for the generation of Linked Data, within the framework of Ontolex-Lemon, and each step of the guidelines serves as a theme for the literature review.

Chapter 3 Methodological Guidelines for Publishing Linked Data

3.1 Introduction

Berners-Lee (2006:n.p.) proposed four Linked Data principles:

1. Use URIs as names for things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL).
4. Include links to other URIs, so that they can discover more things.

However, these principles provide a framework for data once *on* the web, they do not serve as guidelines for publishing Linked Data *to* the web (Wood et al., 2014:234).

In 2011, Villazón-Terrazas et al. proposed methodological guidelines for publishing Linked Data applied to Government Linked Data, which could be applied to any Linked Data project (2011:1). The guidelines took an iterative approach, consisting of of the following main activities (Villazón-Terrazas et al., 2011:4-13):

1. *Specification:* identification and analysis of the data sources; URI design;
2. *Modelling:* identification (and creation) of vocabularies;
3. *Generation:* transformation of the data sources to RDF, data cleansing, inclusion of links to other URIs;
4. *Publication:* publication of the dataset, its metadata, and a Sitemap file; submission thereof to repositories and search engines where appropriate;
5. *Exploitation:* development of applications to exploit the RDF data, providing graphical user interfaces.

Vila-Suero and Gómez-Pérez (2013), Wood et al. (2014:234) and others have published methodological guidelines of their own. However, the language aspect of the RDF data to be published was not taken into account (Vila-Suero et al., 2014:102). In 2014, Vila-Suero et al. proposed methodological guidelines that did consider the language aspect, based on Villazón-Terrazas et al.'s iterative model (2014:103-115):

1. *Specification:* identification and analysis of the data sources; URI design;
2. *Modelling:* identification (and creation) of the domain vocabularies to use; "ontology localization" using one of the following approaches: "(1) multilingual labelling approach, (2) association of the vocabulary to an external lexicon model, and (3) cross-lingual linking or matching approach";
3. *Generation:* transformation of the data sources to RDF, identification of the languages used; consideration of encoding issues;
4. *Linking:* interlinking with external resources at both dataset and instance level;
5. *Publication:* publication of the dataset, its metadata, and a Sitemap file; submission thereof to repositories and search engines, where appropriate.

In 2015, the W3C Best Practices for Multilingual Linked Open Data (BPMLOD) Community Group [15] published guidelines for generating LLD for bilingual and multilingual dictionaries, as well as other lexical resources ("Best practices for …", 2017), and the guidelines for publishing RDF data for a bilingual dictionary are broadly described here: (1) identify the model, (2) select the vocabularies, (3) analyse the data source(s), (4) model the source lexicon, the target lexicon, and the translation set, (5) model a lexical entry, (6) design the URIs, (7) transform the data into RDF, (8) publish the RDF dataset, and (9) publish the metadata (Gracia & Vila-Suero, 2015).

Vila-Suero et al.'s methodological guidelines (which are intended to be iterative) could be subject to further refinement, as demonstrated in this study. For example, the URI design (#1) is dependent on the encoding requirements (#3). Likewise, for Gracia and Vila-Suero's guidelines a proposed improved sequence would be #3, #1, #6, #5, #4.

This chapter will focus on the following:

- the URI design;
- the modelling of a lexicon, a lexical entry, its senses, and translation relations;
- validation of the modelling, by way of visualisation;
- modelling of the metadata, and publishing and generation of the RDF data.

The research questions Q1 and Q2 are addressed in this chapter.

[15] https://www.w3.org/community/bpmlod/

3.2 The URI strategy

A URI has been defined by Archer, Goedertier and Loutas as "a compact sequence of characters that identifies an abstract or physical resource" and the URI "can be further classified as a locator, a name, or both" (2012:4).

The following principles for URIs have been identified:

- URIs should be stable and persistent (Archer, Goedertier & Loutas, 2012:12).
- They should make use of the http:// or https:// scheme (Berners-Lee, 2006; Hogan et al., 2012:23).
- They should be dereferenceable (meaning a representation should be returned) (Heath & Bizer, 2011:10; Hyvönen, 2012:28).
- They should be short (Hogan et al., 2012:25; Archer, Goedertier & Loutas, 2012:12).
- The identifier of a URI should be unique and unambiguous (Simons & Richardson, 2013:85; Keller et al., 2011:28).
- A URI should distinguish between the resource being identified, and the document describing the resource (Van Hooland & Verborgh, 2014:177; Heath & Bizer, 2011:10).
- A URI should be human-friendly (Wood et al., 2014:30-31).

Key concepts relating to the URI are discussed below, followed by discussions of the pattern of the URI and resource identifiers.

3.2.1 Content negotiation

Content negotiation, a mechanism of HTTP, enables for different representations of a resource, depending on the HTTP request of the client (web browser or software agent) (Hyvönen, 2012:26-27). If the client's HTTP header requests a "text/turtle" representation of a resource, and the server configuration allows for this request, the RDF representation of the resource can be generated in the appropriate format and returned to the client; alternatively, the client can be redirected to another address (Hyvönen, 2012:27,57).

3.2.2 Fragment identifiers

Fragment identifiers take the form "#something", and are located at the end of a URI. Although Wood et al. cautioned against the use of fragment identifiers (2014:31), when used to identify sub-resources related to the parent resource, namely the URI to which they are attached, they can be useful. The hierarchical relationship with the URI is clearly indicated by using a fragment identifier, although this hierarchy cannot extend deeper than one level.

Because the URI should resolve "not to the address, but to all known information about the resource" (Sachs & Finin, 2010:1), if an RDF representation of the sub-resource is returned, all information of the parent resource should be returned as well. Likewise, if an RDF representation of the resource is returned, all information of its sub-resources should be returned. This obviates the requirement to have a document describing each sub-resource, as the document describing the parent resource suffices.

Furthermore, the use of fragment identifiers for sub-resources can reduce redundancy when publishing Linked Data, particularly when versioning is considered (see Section 3.8).

3.2.3 The URI pattern

Within the context of this study, six URI use cases have been identified:

U1: A URI that identifies a resource
U2: A URI that identifies a sub-resource to the parent resource
U3: A URI that identifies a version of the resource
U4: A URI that identifies a version of the sub-resource
U5: A URI that identifies a document describing the resource in U1
U6: A URI that identifies a document describing the resource in U3

Archer, Goedertier and Loutas (2012:19) have recommended a pattern for a URI:

```
http://{domain}/{type}/{concept}/{reference}
```

where {domain} is the host, {type} is the resource, for example, id, {concept} refers to a collection or the real world object, and {reference} is the reference for the item being identified (2012:40).

Gracia and Vila-Suero (2015), in guidelines for Linked Data generation for bilingual dictionaries (using the *lemon* model), have recommended the same pattern, with an example URI for a lexical entry given below, where linguistic.linkeddata.es is the host, id is the resource, apertium is the collection, lexiconEN is the source lexicon, and bench-n-en is the reference:

E1: http://linguistic.linkeddata.es/id/apertium/lexiconEN/bench-n-en

Bearing in mind that this is a human interpretation, the following issues are identified:

- id does not provide enough information and can be considered redundant;
- specifically identifying the collection (in this case, apertium) will be problematic when RDF datasets are merged from different collections, resulting in incongruently defined URIs for lexical entries shared between collections;
- the lexicon is identifiable as English, as is the reference (with both containing "en"), making lexiconEN redundant.

A requirement of Ontolex-Lemon (and previously *lemon*, on which the above lexicon and BabelNet were modelled), is that all lexical entries in a lexicon should be of the same language; with translation relations then declared between two lexical entries or two lexical senses, using the *vartrans* module (Cimiano, McCrae & Buitelaar, 2016). If there are two languages, then two lexicons should be defined, one per language. By 2015, BabelNet was supporting 271 languages, thus 271 lexicons, with Flati et al. commenting that *lemon* "forces us to work on a language-by-language basis, whereas in BabelNet this distinction does not need to be made explicit" (2015).

By way of example, the URI for the equivalent lexical entry "bench", in Babelnet's English lexicon, is:

E2: http://babelnet.org/rdf/bench_n_EN

A URI should be agnostic of the selected model. Should the model change, the longevity and persistence of the URIs should not be affected, due to the separation of the URIs and the associated model. By encoding the reference with sufficient information, such as a language code and abbreviated POS with the lemma, as in **E1** (`bench-n-en`) and **E2** (`bench_n_EN`), the URIs are both identifiable as English lexical entries, with POS being 'noun.'

Therefore, **E1** could be simplified as:

`http://linguistic.linkeddata.es/entry/bench-n-en`

And for a lexicon:

`http://linguistic.linkeddata.es/lexicon/en`

This simplified pattern has been applied to each identified use case, with a brief description following.

A URI identifying a resource should be of the form:

U1: `{http(s)://}{Base URI}/{Resource Path}/{Resource ID}`

Where:

- `{http(s)://}` is the `http://` or `https://` scheme
- `{Base URI}` is the namespace, for example, `londisizwe.org`
- `{Resource Path}` is, for example, `entry` for a lexical entry, and `lexicon` for a lexicon
- `{Resource ID}` is the resource identifier, for example, `en-n-abdomen`

A URI identifying a sub-resource to the parent resource should be of the form:

U2: `{http(s)://}{Base URI}/{Resource Path}/{Resource ID}#{Fragment ID}`

Where:

- `{Fragment ID}` is the fragment identifier, for example, `sense1`

The resource identifier is unique relative to the resource path. The fragment identifier is unique relative to the resource identifier.

A URI identifying a version of the resource should be of the form:

U3: `{http(s)://}{Base URI}/{Resource Path}/{Resource ID}/{Version ID}`

Where:

- `{Version ID}` is the version identifier, for example, `2017-09-12`

A URI identifying a version of the sub-resource should be of the form:

U4: `{http(s)://}{Base URI}/{Resource Path}/{Resource ID}/{Version ID}`
`#{Fragment ID}`

When implementing versions of a resource, each version should be accessible, with the version identifier unique to the resource identifier. **U1** should always resolve to the latest version (Archer et al., 2012:6).

A URI identifying a document describing the resource in **U1** (or **U3**) should be of the form:

U5: `{http(s)://}{Base URI}/{Document}/{Resource Path}/{Resource ID}`
U6: `{http(s)://}{Base URI}/{Document}/{Resource Path}/{Resource ID}/`
`{Version ID}`

Where:

- `{Document}` could refer to `rdf` which serves as the RDF representation (using any serialisation). If content negotiation is in place, then `page` could refer to the HTML page.

A document describing **U2** (or **U4**) should not be required, and instead it should resolve to **U5** (or **U6**).

3.2.4 Resource identifiers

In the literature, repeated references are made to the human-friendliness requirement of URIs, such as they should be "human readable" (Hogan et al., 2012:25), "user-friendly" (Archer, Goedertier & Loutas, 2012:18), "meaningful" (Villazón-Terrazas et al., 2011:6), and they should use "natural keys" (Wood et al., 2014:30-31; Heath & Bizer, 2011:43). Defined as "meaningful URIs" by Vila-Suero et al. (2014:106) and "descriptive URIs" by Labra Gayo, Kontokostas and Auer, these URIs are typically used "with terms in English or in other Latin-based languages" (Labra Gayo, Kontokostas & Auer, 2013:5).

"Opaque URIs" are defined by Labra Gayo, Kontokostas and Auer as "resource identifiers which are not intended to represent terms in a natural language" (2013:6) and both Labra Gayo, Kontokostas and Auer (2013:6) and Vila-Suero et al. (2014:107) suggest that their use in a multilingual context is preferable to avoid language bias. Vila-Suero et al. argue that doing so within the Semantic Web context is acceptable, as "resource identifiers are intended for machine consumption so that there is no need for them to be human readable" (2014:107). This view may be accurate within the larger context of the Semantic Web where data models are largely language-agnostic (Ehrmann, 2014:402), but opposes a principle of Linked Data that a URI can be looked up by both a web browser for human consumption and a software agent (Hyvönen, 2012:26).

Because this study is localised in South Africa, using Roman alphabet-based languages, pragmatism supports a descriptive URI approach using English, although the senses and concepts could be modelled using opaque URIs, as done by Babelnet (Flati et al., 2015).

The resource identifier for lexical entries will take a similar form to **E1** (including the resource path):

```
{Resource Path}/{Language Code}-{POS}-{Lemma}
```

Where:

- `{Resource Path}` will be `entry`, expressed in English

- `{Language Code}` will be the shortened language code, in lowercase, applicable to the lexical entry, as defined by ISO 639-1, and if none available, then ISO 639-2 or ISO 639-3 (de Melo, 2015:2)
- `{POS}` will be an abbreviated form of the POS, expressed in English
- `{Lemma}` will be the lemma, in lowercase, replacing hyphens and spaces with underscores

In Ontolex-Lemon, a lexical entry can have only one POS, and as there is the constraint that it cannot be combined with other lexical entries of different languages in one lexicon (Cimiano, McCrae & Buitelaar, 2016), it is best to include both the language and the POS of the lexical entry in the identifier so as to avoid conflicts with other lexical entries which share the same lemma, for example:

isiXhosa: `xh-n-isibindi`
isiZulu: `zu-n-isibindi`

For a lexicon, the resource identifier will take the form:

```
{Resource Path}/{Language Code}
```

Where:

- `{Resource Path}` will be `lexicon`, expressed in English
- `{Language Code}` will be the shortened language code, in lowercase, of the lexical entries

The resource identifier, when combined with the resource path, will sufficiently identify the lexical entry or lexicon to allow for the representation of any language. Lemmas containing characters not in the Roman alphabet is addressed in Chapter 5.

A notable feature of the core Ontolex-Lemon model is that directionality from source to target language is not preserved. Although both Labra Gayo, Kontokostas and Auer (2013:6) and Vila-Suero et al. (2014:107) talk of language bias, this is a human interpretation only; within the framework of the model, each lexicon is treated equally, with a human user or software agent able to access a lexicon and its translation

equivalents from any direction, with the identified ontology entity serving as the connector between the entries. For more nuanced translation equivalence, however, directionality does become important.

3.3 The description of resources

One of the principles of Linked Data is that URIs should be dereferenceable, and when information is returned to the human user or the software agent, it should provide an RDF description of "all known information about the resource" (Sachs & Finin, 2010:1). The information returned should not just be limited to RDF triples that describe the resource using literals or by linking to other resources, but RDF triples describing the following information should be considered for inclusion:

- related resources;
- resource metadata, such as provenance;
- the dataset containing the resource (Heath & Bizer, 2011:45).

When resolving a URI which describes a lexicon, for example, `http://londisizwe.org/lexicon/en`, the potential size of the lexicon may prevent the inclusion of all information of its related resources, in this case the lexical entries, unless by data dump. However, for a lexical entry, for example, `http://londisizwe.org/entry/en-n-abdomen`, this is possible, and the following additional information should be included:

- document description for the the lexical entry;
- lexical entry metadata;
- lexical entry provenance information;
- identification of the lexicon to which the entry belongs;
- brief description of other lexical entries, resources and ontology entities related to the lexical entry.

3.4 Modelling a lexical entry

In Chapter 1, two lexical entries were introduced, namely *abdomen* and *isisu*. *abdomen* is used below to demonstrate the modelling of a lexical entry using Ontolex-Lemon. Both *abdomen* and *isisu* will be used to demonstrate the modelling of translation relations.

The resource identifier for *abdomen* is: `en-n-abdomen`. The URIs associated with `en-n-abdomen` are as follows:

U1: `https://londisizwe.org/entry/en-n-abdomen`

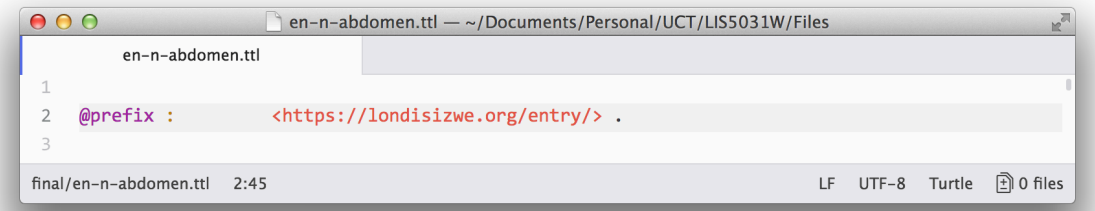
U5: `https://londisizwe.org/rdf/entry/en-n-abdomen`

In natural language, `en-n-abdomen` can be described as:

- It is a single word.
- It is a noun.
- It is an English term, therefore it is in the English lexicon.
- The lemma is "abdomen".
- It is denoted by the concept: Abdomen[16] on Wikipedia.
- It has a translation equivalent in isiXhosa: `xh-n-isisu`.

Using the guidelines provided by Ontolex-Lemon[17], modelling of the lexical entry and the additional information identified in Section 3.3, in Turtle RDF syntax, is as follows:

3.4.1 Defining the namespace



Code Example 3-1: Defining the namespace

[16] https://en.wikipedia.org/wiki/Abdomen
[17] https://www.w3.org/community/ontolex/wiki/Final_Model_Specification

Where:

- Line 2: defines the namespace relative to the resource being described.

3.4.2 Describing the lexical entry: `en-n-abdomen`

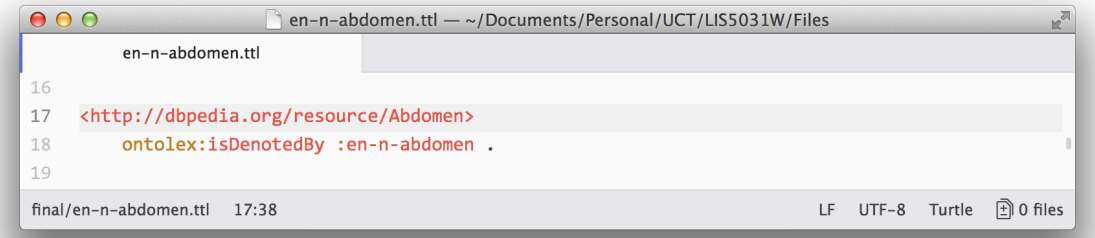
Code Example 3-2: Describing the lexical entry en-n-abdomen

Where:

- Line 5: the resource is defined as an instance of the `ontolex:LexicalEntry` class, and `Word` is a sub-class of `LexicalEntry`.
- Line 6: the lexical entry's POS is defined as a noun, using the LexInfo vocabulary.
- Lines 7-8: the language of the lexical entry is defined, using the DCMI Metadata Terms language property, referencing both the Library of Congress vocabulary for the representation of language names and the Lexvo ontology.
- Line 9: the `canonicalForm` property relates the lexical entry to its canonical form, and line 15 indicates the lemma of the lexical entry.
- Line 10: in addition to the `canonicalForm` property, use of the `rdfs:label` property is recommended for RDFS-based systems expecting a RDFS label (Cimiano, McCrae & Buitelaar, 2016).
- Lines 10 & 15: both plain literals are language-tagged strings.

- Line 11: as there are not any senses associated with this lexical entry, the lexical entry is mapped to an ontology entity which represents the meaning thereof, using the `denotes` property (Cimiano, McCrae & Buitelaar, 2016). The ontology used is DBpedia, the RDF equivalent of Wikipedia.

3.4.3 Describing the ontology entity related to the lexical entry

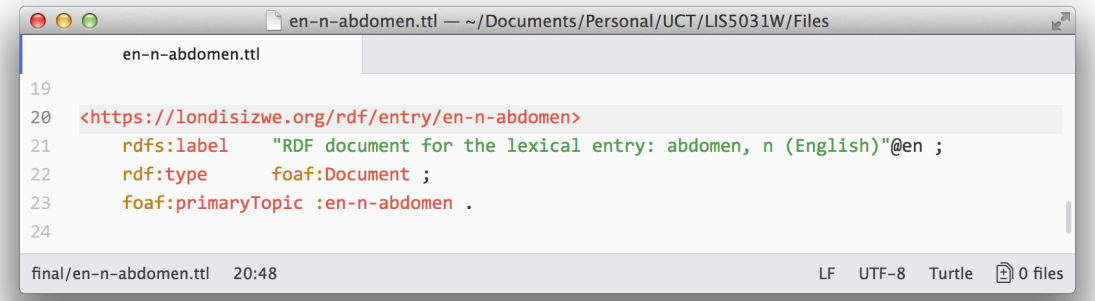


Code Example 3-3: Describing the ontology entity

Where:

- Lines 17-18: the ontology entity is described, using the inverse of the `denotes` property.

3.4.4 Describing the document



Code Example 3-4: Describing the document

Where:

- Lines 20-23: the document is described, with `foaf:primaryTopic` identifying the lexical entry as the topic of the document.

3.4.5 Identified limitations

It does not appear possible, at time of writing, to model the lexical entry's relationship to the lexicon using Ontolex-Lemon; this is unlike the RDF/OWL version of LMF which has the property `isPartOf` (McCrae et al., 2012:705), although it can be modelled using the `dct:isPartOf` property. However, according to McCrae et al., "the necessity to have all words grouped into a lexicon is no longer core, but remains a useful feature" (2017:590).

3.5 Modelling the lexicon

The resource identifier for the English lexicon is: `en`. The URIs associated with `en` are:

U1: `https://londisizwe.org/lexicon/en`

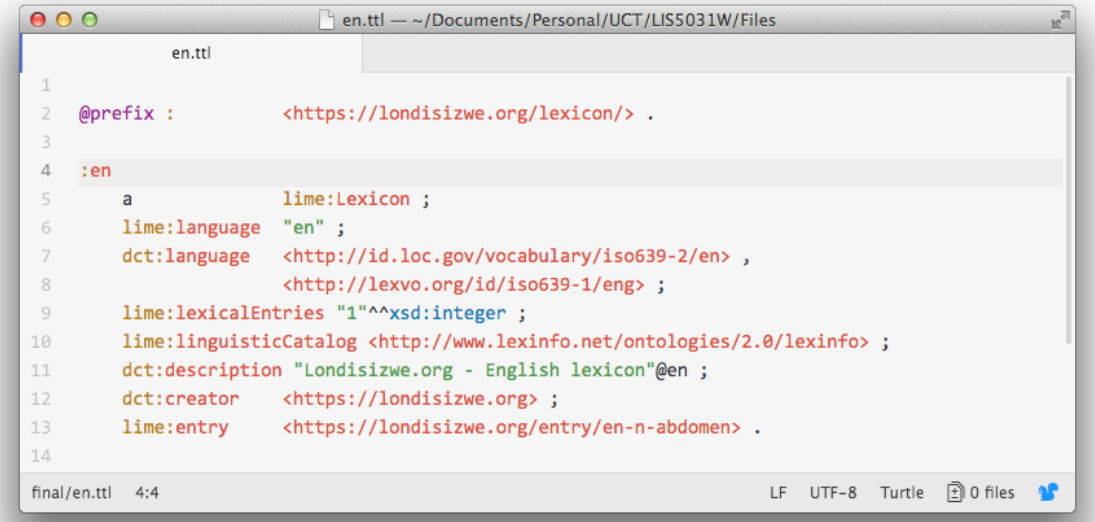
U5: `https://londisizwe.org/rdf/lexicon/en`

In natural language, `en` can be described as:

- It is a lexicon.
- It contains one lexical entry: `en-n-abdomen`
- It only contains English lexical entries, therefore it is an English lexicon.

Using the guidelines provided by Ontolex-Lemon, the lexicon is modelled in Section 3.5.1.

3.5.1 Describing the lexicon: en

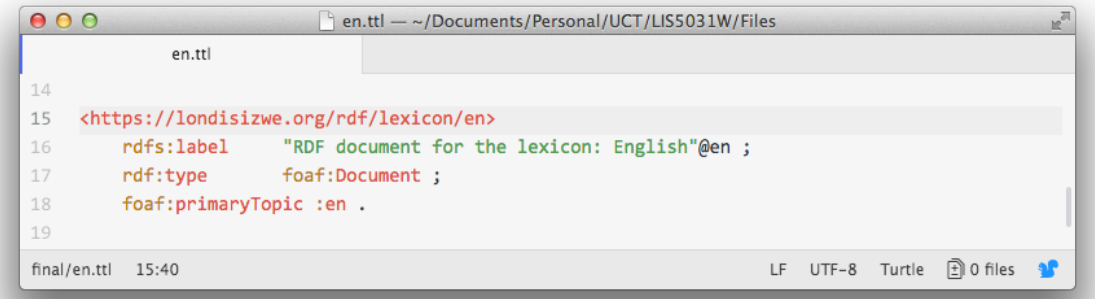


Code Example 3-5: Describing the lexicon en

Where:

- Line 5: the resource is defined as a lexicon.
- Line 6: this indicates the language of the lexicon. For "en", the expected value is a (untagged) literal.
- Line 9: the number of entries in the lexicon is indicated here.
- Line 10: the POS for the lexical entry was described using the Lexinfo vocabulary, so the linguistic catalogue used is defined here.
- Line 13: represents the collection of lexical entries, in this instance, there is only one.

3.5.2 Describing the document



Code Example 3-6: Describing the document

3.6 Translation equivalence in dictionaries

As both `en-n-abdomen` and `xh-n-isisu` denote the same ontology entity, they share the same interpretation, thus their equivalence as lexical entries can be inferred (Cimiano, McCrae & Buitelaar, 2016). If a third lexical entry, with the identifier `af-n-buik` (for an eventual equivalent lexical entry in Afrikaans), had to denote the same ontology entity, then it too would share equivalence with `en-n-abdomen` and `xh-n-isisu`, thus allowing for easy extensibility to multilingualism. However, if one reviews the definitions provided by resources for the lexical entries *abdomen* and *isisu*, the following is observed:

3.6.1 BabelNet

A multilingual encyclopaedic dictionary and semantic network, BabelNet groups synonyms which express "a given meaning" in a range of different languages together as a set, called a Babel synset (BabelNet, n.d.). For the Babel synset containing the term *abdomen*, ("abdomen • stomach • belly …", n.d.) the following English synonyms are shown:

en abdomen, stomach, belly, venter

3.6.2 Oxford English Xhosa Dictionary

Using this bilingual dictionary, a unidirectional dictionary from English to isiXhosa ("abdomen", 2013:1; "belly", 2013:52; "stomach", 2013:626), the articles for *abdomen*, *belly*, and *stomach* are as follows:

> **abdomen,** *n*. isisu.
> **belly,** *n.* (colloq.) isisu.
> **stomach,** *n.* isisu.

3.6.3 The Greater Dictionary of Xhosa

In this multilingual dictionary, a unidirectional dictionary from isiXhosa to English and Afrikaans (Pahl, Pienaar & Ndungane, 1989:225), the article for *isisu* is as follows:

> isí**·sù**:
> **1** [English] stomach, rumen
> [Afrikaans] maag, pens (organ)
>
> **2** [English] abdomen, stomach, soft underbody (as seen from outside)
> [Afrikaans] die abdomen soos van buite gesien

3.6.4 EXDN

Using EXDN ("Abdomen", 1935:1; "Stomach", 1935:91), the articles for *abdomen* and *stomach* are as follows:

> **Abdomen.** Isisu.
> **Stomach.** Uluusu lomntu.

3.6.5 DBpedia

On DBpedia, *abdomen* and *stomach* are defined as two separate ontology entities:

> **Abdomen** http://dbpedia.org/resource/Abdomen, where *Abdomen* is an abdominal cavity comprising several organs.
> **Stomach** http://dbpedia.org/resource/Stomach, where *Stomach* is an organ within the abdominal cavity.

This shows that ontological equivalence is not always shared between lexical entries of different languages. According to Hirst (2014:7), one "cannot rely on an ontology as an interlingual representation or as a nonlinguistic representation for inference; there is, in practice, no clean separation between the conceptual and the linguistic."

Whilst the denotation may hold true for a unidirectional lexical entry, assuming the same denotation in the reverse direction may introduce inaccuracies. It would therefore be preferable to indicate ontological equivalence at sense level. In a printed or electronic dictionary, a sense is defined as a word having "several distinct meanings or 'word senses'" (Atkins & Rundell, 2008:264). In an ontology-lexicon, a sense is defined as a reification "between a lexical entry and the concept it *evokes*", whilst simultaneously providing a hook into the ontology denoted in the lexical entry, enabling the context of the concept to be explored further (Cimiano et al., 2012:3). By way of example: `en-n-belly` evokes the concept "stomach", denoted by the ontology entity `dbr:stomach`; however, the sense indicates the condition in which the concept `dbr:stomach` can be evoked; "belly" would be inappropriate in a more formal register. According to Cimiano et al. (2012:3), a sense has three facets:

1. a reification of a lexical entry and the concept denoted by the ontology entity, for example, it links the lexical entry `en-n-belly` to the concept "stomach";
2. a subset of the lexical entry, within the context of the evoked concept, for example, `en-n-belly` could only be used colloquially to describe the concept "stomach"; and
3. a concept of the evoked concept, and if added to the ontology, it would be a sub-class, for example, "belly" would be a sub-class of "stomach".

To ensure accurate representation, a minimum of one sense should be modelled for each lexical entry. When modelling translation relations at sense level, the *vartrans* module of Ontolex-Lemon provides several options:

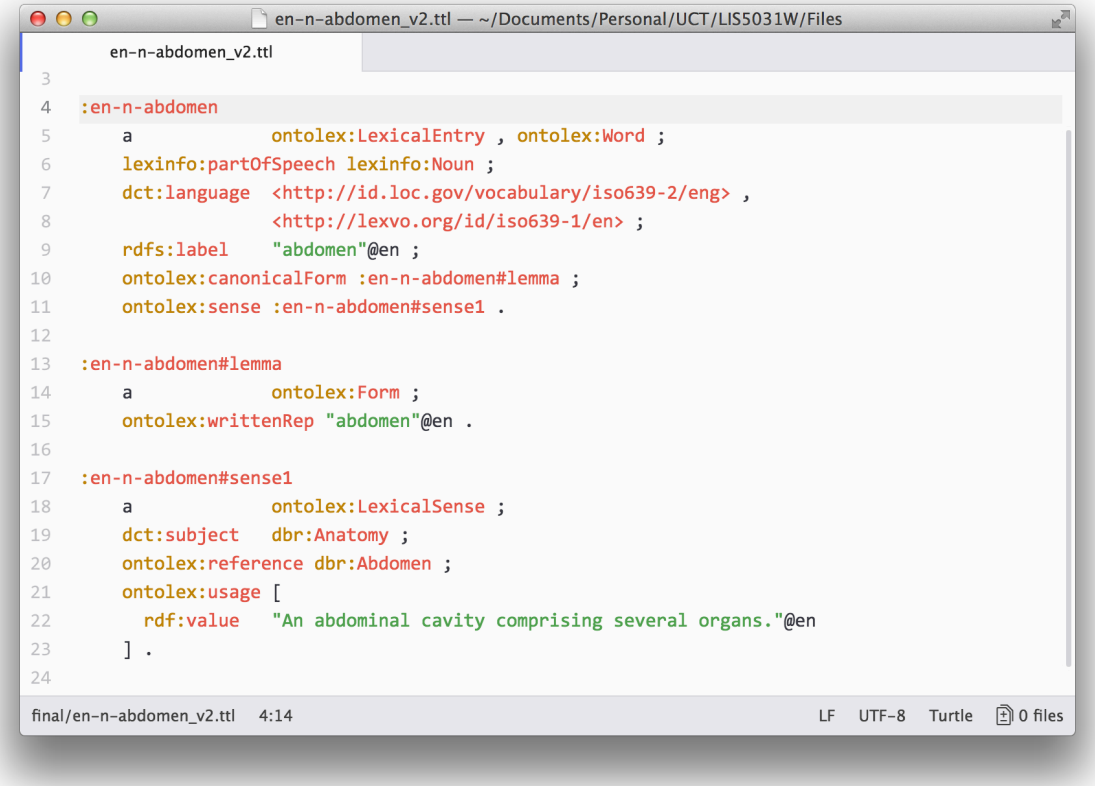
- *Use Case 1:* Translation as shared reference
- *Use Case 2:* Translation as a relation between lexical senses
- *Use Case 3:* Translatable As

According to Gouws and Prinsloo, "translation equivalents will be determined by the context and the cotext of the source language item" and translation equivalents should not be "isolated from their typical contexts and cotexts" (2005:153). With this in mind, in the next section, the translation equivalence use cases provided for by the *vartrans* module are modelled.

3.7 Modelling translation relations

Continuing with `en-n-abdomen` and `xh-n-isisu`, a third lexical entry, `af-n-boep`, is introduced. Modelling of the lexical entries is as follows:

3.7.1 Describing the lexical entry: `en-n-abdomen`



Code Example 3-7: Describing the lexical entry en-n-abdomen

Where:

- Lines 11, 17-23: the sense is defined and described.
- Lines 21-23: this lexicographic definition serves as a comment on semantics, and is language-tagged.

3.7.2 Describing the lexical entry: `xh-n-isisu`

Code Example 3-8: Describing the lexical entry xh-n-isisu

Where:

- Lines 9, 15-40: three distinct senses are defined for this lexical entry; indicating that the word is polysemous (Gouws & Prinsloo, 2005:151).
- Lines 19-22, 28-31, 37-40: each serves as a comment on semantics. In the absence of a definition, the latter is a cotextualisation (Gouws & Prinsloo, 2005:127), indicated by the quotation marks.
- Line 39: "(Informal)" indicates context of use, and is a stylistic label (Gouws & Prinsloo, 2005:127-130).

3.7.3 Describing the lexical entry: `af-n-boep`

af-n-boep.ttl — ~/Documents/Personal/UCT/LIS5031W/Files

af-n-boep.ttl

```
1
2  :af-n-boep
3      a              ontolex:LexicalEntry , ontolex:Word ;
4      lexinfo:partOfSpeech lexinfo:Noun ;
5      dct:language   <http://id.loc.gov/vocabulary/iso639-2/afr> ,
6                     <http://lexvo.org/id/iso639-1/af> ;
7      rdfs:label     "boep"@af ;
8      ontolex:canonicalForm :af-n-boep#lemma ;
9      ontolex:sense :af-n-boep#sense1 .
10
11 :af-n-boep#lemma
12     a              ontolex:Form ;
13     ontolex:writtenRep "boep"@af .
14
15 :af-n-boep#sense1
16     a              ontolex:LexicalSense ;
17     dct:subject    dbr:Anatomy ;
18     ontolex:reference dbr:Stomach ;
19     ontolex:usage [
20       rdf:value    "(Informeel) 'Sy boep is groot van al die bier wat hy drink.'"@af ,
21       rdf:value    "(Engafanelekanga) 'Isisu sakhe sikhulu kuku sela kakhulu ibhiye.'"@xh ,
22       rdf:value    "(Informal) 'His belly is big from too much beer.'"@en
23     ] .
24
```

final/af-n-boep.ttl 2:11 LF UTF-8 Turtle 0 files

Code Example 3-9: Describing the lexical entry af-n-boep

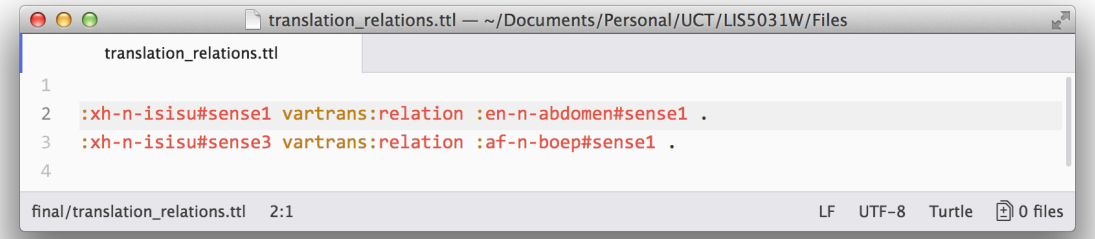
Use Case 1: Translation as shared reference

For this use case, lexical senses of two lexical entries in different languages can be expressed as ontologically equivalent by pointing to the same ontology entity, thus "they are clearly translations as they have the same interpretation" (Cimiano, McCrae & Buitelaar, 2016).

- `:en-n-abdomen#sense1` and `:xh-n-isisu#sense1` share full ontological equivalence (indicated by the `ontolex:usage` property) (Gouws & Prinsloo, 2005:154).
- `:xh-n-isisu#sense2` and `:af-n-boep#sense1`, despite being ontologically equivalent, share only partial equivalence (Gouws & Prinsloo, 2005:155).

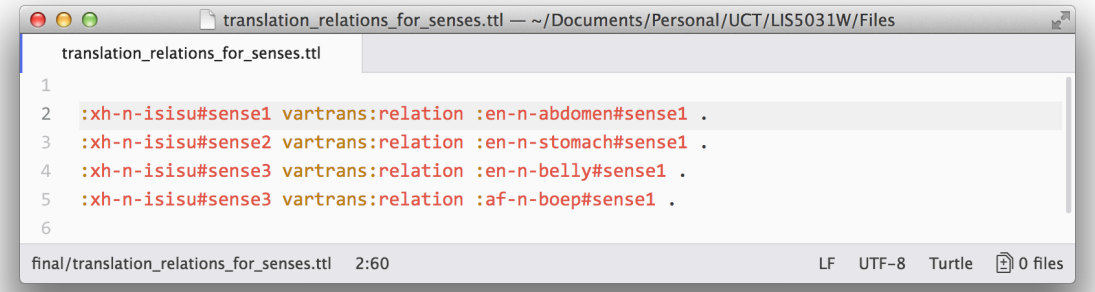
Use Case 2: **Translation as a relation between lexical senses**

For this use case, the ontology entities for two senses may differ, but direct equivalence can still be expressed using the Translation class, which is defined as: "a sense relation expressing that two lexical senses corresponding to two lexical entries in different languages can be translated to each other without any major meaning shifts" and the translation relation between two lexical senses is reified with this class (Cimiano, McCrae & Buitelaar, 2016). The translation property is also available as a shorthand method to express the translation relation between the two senses (Cimiano, McCrae & Buitelaar, 2016), as in Code Example 3-10 for `:xh-n-isisu`.



Code Example 3-10: Describing translation relations using the shorthand method

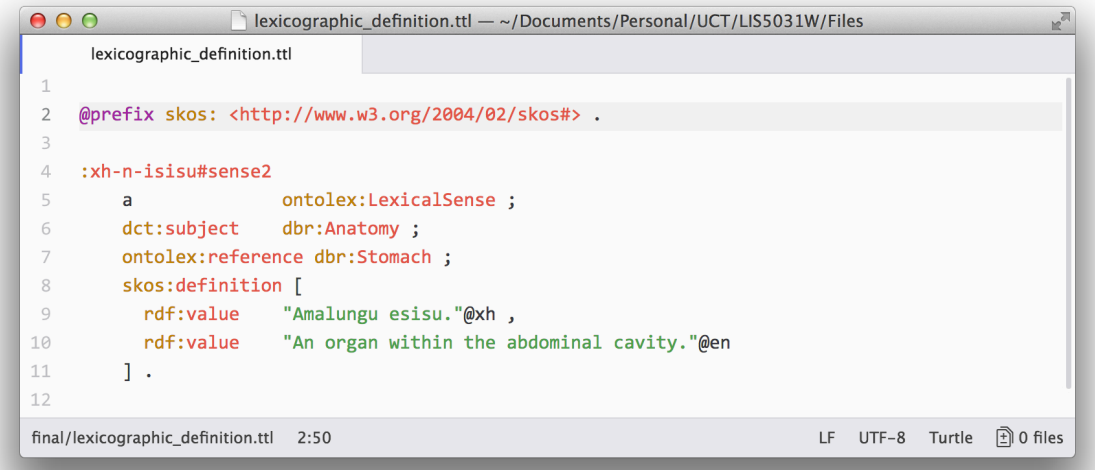
A polysemous word may not have a translation equivalent in the target language with the same polysemous senses. Instead, a translation equivalent may have to be provided for each one of the senses (Gouws and Prinsloo, 2005:151-152), as modelled in Code Example 3-11.



Code Example 3-11: Modelling translation equivalents for senses

Equivalents like *abdomen*, *stomach* and *belly* represent some of the senses of *isusu*. According to Gouws and Prinsloo, "from a semantic perspective it would be wrong to argue that anyone [sic] of these equivalents can be regarded as the meaning of the word" *isisu*; the isiXhosa word *isisu* does not mean *belly* "but in a specific environment it could be translated with the word" *belly* (2005:153).

As has been shown, when representing translation equivalents in a bilingual or multilingual environment, it is of "extreme importance" to include cotexts and typical contexts applicable to a sense in the source language, instead of relying solely on an ontology entity to represent the specific meaning (Gouws & Prinsloo, 2005:153).



Code Example 3-12: Modelling a lexicographic definition

The emphasis placed on context and cotexts leads to the following limitations identified for the model:

1. Cotextualisation at sense level cannot be explicitly specified.
2. Context at sense level cannot be explicitly specified.

However, a lexicographic definition can be explicitly specified by replacing the `ontolex:usage` property with `skos:definition`, as shown in Code Example 3-12.

Gracia et al. (2018:7) discuss the construction of a bilingual dictionary using one or more pivot languages to create new inferred translation pairs, for example:

English to isiXhosa: *belly* → *isisu*
isiXhosa to Afrikaans: *isisu* → *boep*

with the inferred translation pair:

English to Afrikaans: *belly* → *boep*

Inaccurately defining context, cotextualisations and a lexicographic definition can result in ambiguities, resulting in possible meaning shifts. Furthermore, the provenance of inferred equivalences should be identified as such, with a label indicating that it is an inferred equivalent.

Using the class Translation Set, one and more sense translations can be grouped into sets indicating a shared commonality, such as derivation from the same language resource, or that they were "`"Automatically translated"@en`" (Cimiano, McCrae & Buitelaar, 2016). While modelling provenance would be better done at the lexical entry level, this class may serve as a solution to model the metadata of a language pair.

Use Case 3: **Translatable As**

Underspecification is defined by Saint-Dizier as a "way of representing, with a certain degree of generality, the semantics of lexical items whose meaning remains generic or vague and may vary depending on other constituents in the sentence" (2000:113). The property `translatableAs` serves as a way to represent underspecification (Cimiano,

McCrae & Buitelaar, 2016). Although none of the previously introduced lexical entries demonstrate this use case, it is included here for informational purposes.

The translation relations between the senses of the lexical entries in three lexicons is visualised in Figure 3-1. A prototypical lexical entry is visualised in Figure 3-2.

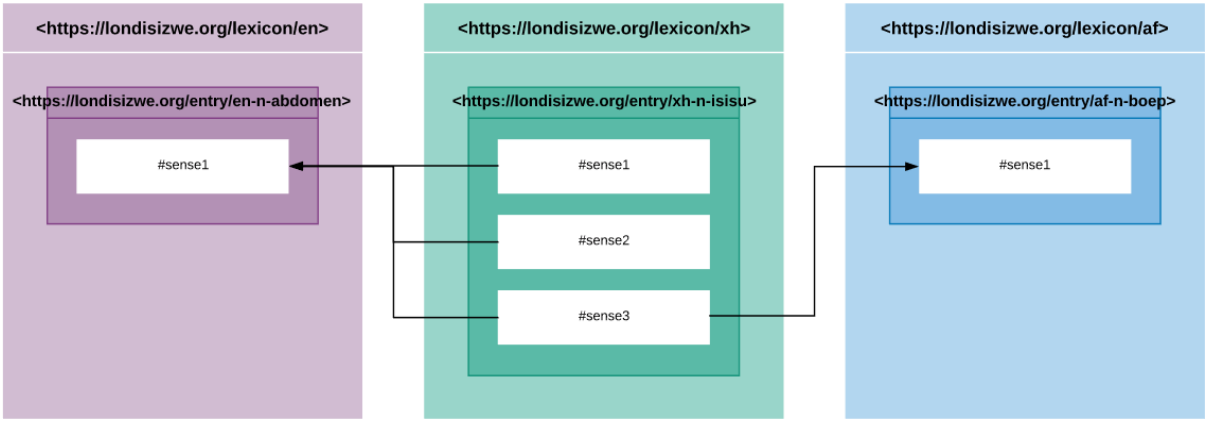


Figure 3-1: Visualisation of the translation relations for en-n-abdomen, xh-n-isisu *&* af-n-boep

Using the *vartrans* module, the source and target language pair is set, but this is unidirectional only. For bidirectionality, a second triple would need to set the target language to the source, and the source to the target. For multilingual resources, this would require modelling a translation relation between every equivalent sense.

The lexical entry for `en-n-abdomen` is visualised in Figure 3.2.

Figure 3-2: Visualisation of the lexical entry en-n-abdomen

3.8 Modelling provenance and versioning

Knowledge is "partial/incomplete/imperfect, with very few exceptions" (Di Maio, 2015:4). The very nature of Linked Data is about its relationships and in the context of LLD, different datasets can be interlinked, thus allowing an existing lexicon to be extended; a powerful notion for under-resourced languages (Berners-Lee, 2009; McCrae

et al., 2012:702). When one considers the retrodigitisation of language resources, according to Bouda and Cysouw, the challenge is not the encoding, but "the continuing update, refinement, and interpretation" of the digitised data, ensuring traceability as the dataset changes (2012:16).

Ontologies, vocabularies and RDF datasets evolve over time (Hyvönen, 2012:94). Change results from the correction of errors, change to the underlying model by the addition of new concepts and properties, and changes out in the world and to our understanding of it (Hyvönen, 2012:95). An RDF dataset typically comprises two levels of data: primary and metadata (Hyvönen, 2012:103); for a language resource it is comprised of three levels of data (primary level, metadata level, and linguistic annotations) (Eckart, Riester & Schweitzer, 2012:70), and for a language resource representing multiple languages, the researcher has identified four levels of RDF triples:

Primary level: lexicons, lexical entries and senses
Secondary level: translation relations
Annotation level: annotations
Metadata level: provenance, licensing and versioning
(Gracia et al., 2018:3; Heath & Bizer, 2011:52-53).

The four levels are visualised in Figure 3-3.

At the primary level, within the framework of Ontolex-Lemon, data consists of the lexicons, the lexical entries contained in each lexicon and the senses contained by each lexical entry. At the secondary level, each translation relation is a language pair in its own container, connecting the senses across lexical entries and lexicons. Annotations can then be made for lexical entries, their senses, and translation relations. The metadata would describe the data at all of the previous levels, as well as the versioning.

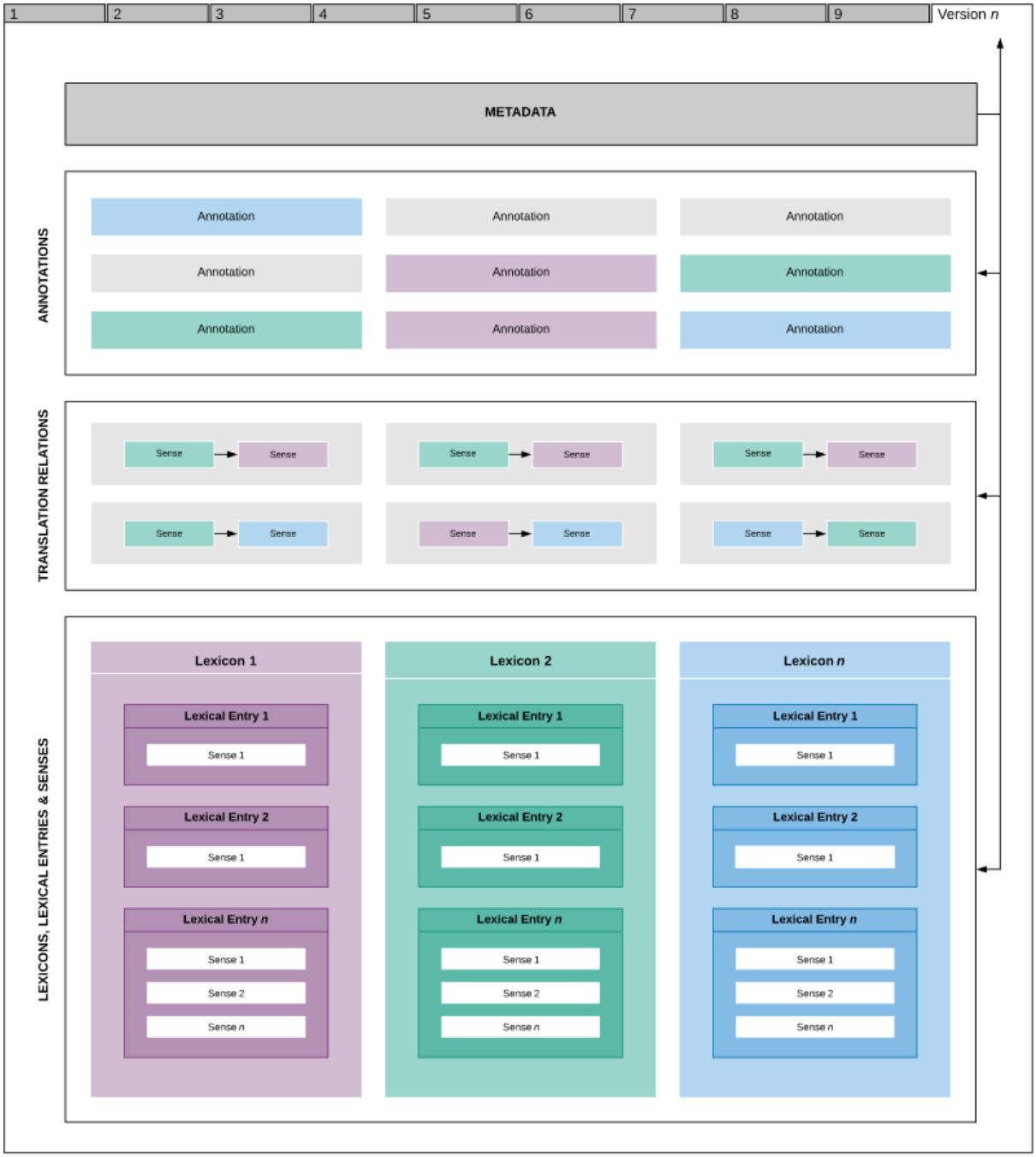


Figure 3-3: The four levels of data, with versioning

3.8.1 Versioning

Since 2015, Babelnet has employed versioning for their BabelNet-lemon schema description, although this is applied globally "for RDF, currently 3.0" with Flati et al. acknowledging that "maybe a more sophisticated infrastructure would be needed in order to express more complex versioning description needs" (2015). Although Gracia et

al. (2018:5-7) mention metadata in the generation and publication of RDF data of the Apertium Bilingual Dictionaries, versioning of the RDF data was omitted. Other than a cursory mention by McCrae et al. (2012:711), Gracia et al. (2018:3), Eckart, Riester and Schweitzer (2012:66), van Erp (2012:63) and De Rooij et al. (2016:198), the concept of versioning does not appear to have been explored further within the LLD domain, with Flati et al., when referring to the available vocabularies, saying that heavy changes are not accounted for, "and this aspect might thus be investigated in more detail in the [near] future by the whole community" (2015).

Gracia et al. (2018:5), referring to the generation of RDF, described the generation of one file for each of their lexicons and a third file for the translations. If versioning had to be implemented, it would possibly be by versioning each lexicon file, analogous to BabelNet's schema description versioning. However, this approach could quickly become unmanageable and versioning would perhaps be better implemented at the lexical entry level with a file representing a lexical entry, its senses, and the translation relations for which any one of its senses is the source.

Three components to versioning have been identified:

- versioned URIs for lexicons, lexical entries and senses;
- provenance metadata, with the latest version mapping to previous versions (Van Erp, 2012:63), and
- the generation of a file for each version of the lexical entries and lexicons.

Versioning of URIs was introduced with the URI use cases: **U3** and **U4** in Section 3.2.3, so only provenance metadata and the generation of files are discussed here.

3.8.2 Modelling provenance for a lexical entry, its senses, and translation relations

Provenance is defined by the W3C Provenance Working Group (W3C, 2013b) as "a record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering a piece of data or a thing."

When data are reused, either by interlinking or by working with the downloaded RDF dataset, trust in the data, as well as the repository supplying the data, is a major factor

contributing to its reuse (Faniel & Yakel, 2017:110). In an open environment such as the web, provenance, using PROV-O, DCMI Metadata terms and versioned URIs, can serve as a trust marker by providing a systematic schema for the documentation of the data (Faniel & Yakel, 2017:112; W3C, 2013b; Tennis, 2007:86; Flati et al., 2015).

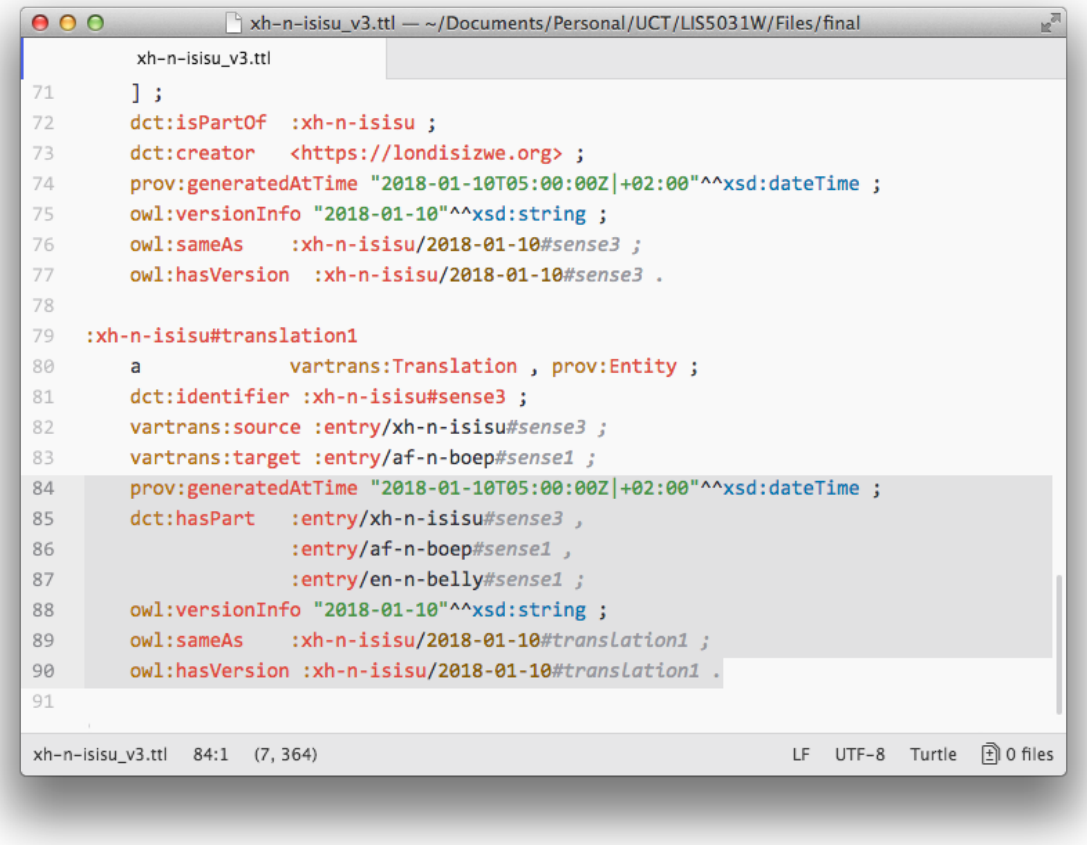
The following suggested metadata can be described in RDF:

- Identify each lexical entry, sense, and translation relation as a `prov:Entity`.
- Record the `prov:generatedAtTime` property for each.
- Include the date a lexical entry, sense or translation relation was last changed, using `dct:modified`.
- Identify the person or organisation responsible for creating the lexical entry or sense, using `dct:creator`.
- Identify the source from which a lexical entry was primarily derived, using the `prov:hadPrimarySource` property.
- Identify the sources from which a lexical entry, sense or translation relation was derived, using the `dct:source` property.
- Identify one or more contributors for a lexical entry, sense or translation relation using `dct:contributor`. This can be a person, organisation or service.
- For a lexical entry, use `dct:license` to point to an applicable Creative Commons[18] licence.
- For a lexical entry or sense, use `dct:isPartOf` to denote inclusion in a lexicon or a lexical entry, respectively.
- For a translation relation, use `dct:hasPart` to denote both the source and target language.
- For a lexical entry, use `owl:sameAs` to indicate that **U1** is the same as the latest version of **U3**.
- For a sense or translation relation, use `owl:sameAs` to indicate that **U2** is the same as the latest version of **U4**.
- For a lexical entry, sense or translation relation, indicate the version using `owl:versionInfo`.

[18] https://creativecommons.org/

- For a lexical entry, sense or translation relation, use `dct:hasVersion` to show the previously generated versions, using the versioned URIs (**U3** for lexical entries and **U4** for senses and translation relations).

Modelling of the lexical entry `xh-n-isisu`, its senses, a translation relation, and the associated metadata is shown in Code Examples 3-13 and 3-14.



Code Example 3-13: Modelling of a translation relation of a sense of the lexical entry, xh-n-isisu*, and its associated metadata*

Where:

- Lines 84-90: indicate the triples for the metadata of a translation relation.

Code Example 3-14: Modelling of the lexical entry xh-n-isisu *and its associated metadata*

Where:

- Lines 12-21: indicate the triples for the metadata of the core lexical entry.
- Lines 36-43: indicate the triples for the metadata of a sense.

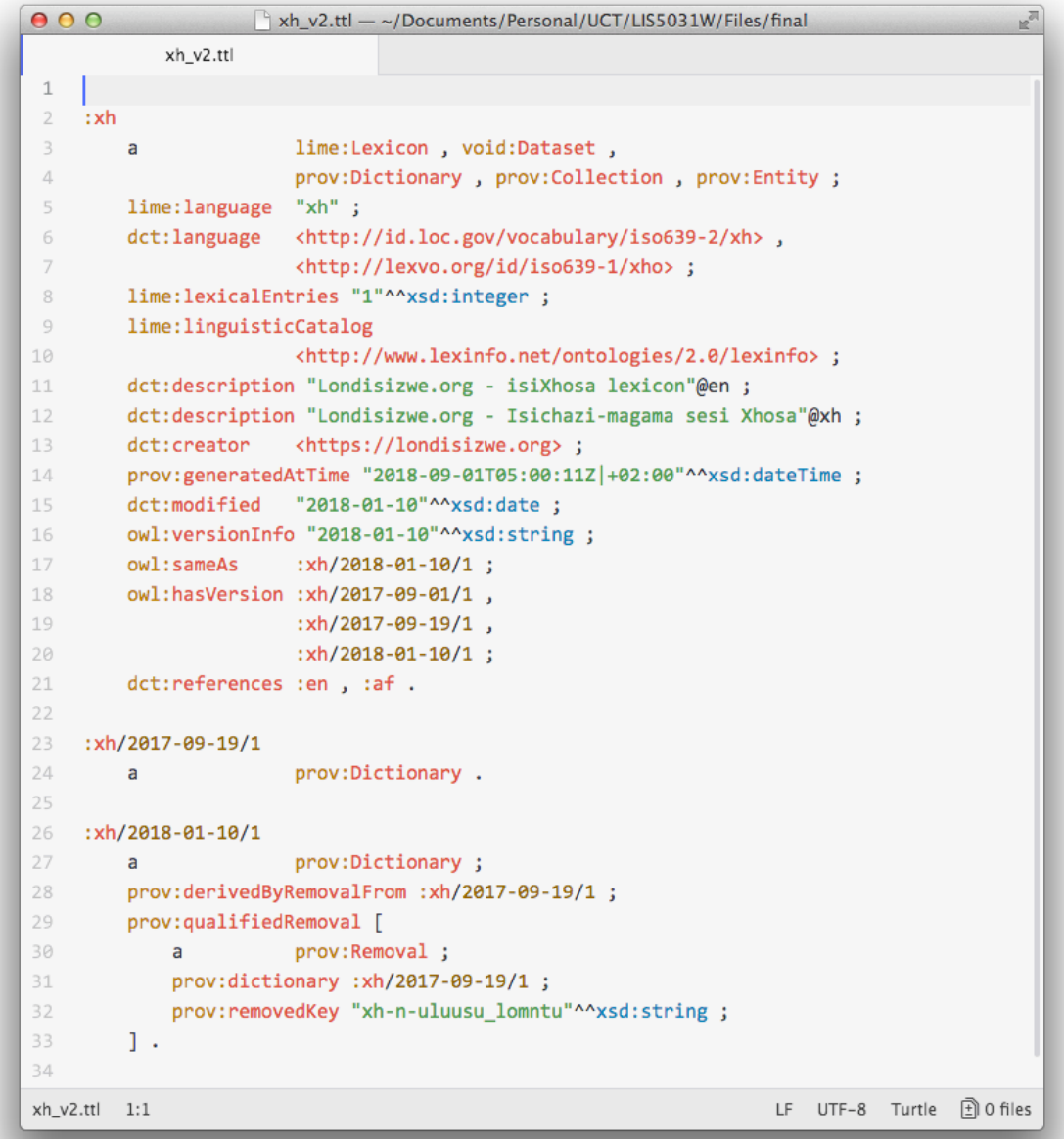
3.8.3 Modelling provenance for a lexicon

In Section 3.5.1, the lexicon was described using *lime*. The suggested additional metadata can be described in RDF:

- Identify each lexicon as a `prov:Entity` and a `void:Dataset`.
- Record the `prov:generatedAtTime` property for each.
- Include the date a lexicon was last changed, using `dct:modified`.
- Indicate other lexicons within the same namespace using `dct:references`.
- Use `owl:sameAs` to indicate that **U1** is the same as the latest version of **U3**.
- Indicate the version using `owl:versionInfo`.
- Use `dct:hasVersion` to show the previously generated versions.

However, the metadata above only describes the lexicon; it does not describe when a lexical entry was inserted into (or removed from) said lexicon. W3C Provenance Working Group has issued a Note on PROV-Dictionary, and it "introduces a specific type of collection, consisting of key-entity pairs", and it allows for PROV-O to be extended, thereby expressing the members of the lexicon as well (W3C, 2013a).

Version three of a lexicon is modelled in Code Example 3-15.



Code Example 3-15: Modelling version three of a lexicon

Where[19]:

- Line 23-24: declares the previous version to be a dictionary. The previous two dictionary entries would have been listed in the file of the previously published URIs: https://londisizwe.org/lexicon/xh/2017-09-19/1

[19] For Line 32, the identifier "xh-n-uluusu_lomntu" was generated according to the original lexicon. However, the spelling of "ulusu" was incorrect in the original lexicon. Regardless, its retention as an identifier will not impact any subsequent changes to spelling in updated entries.

- Line 26-33: the current version is declared to be a dictionary.
- Line 28: states that the current version was derived from the previous version, by means of removing keys.
- Lines 29-33: indicates the key that was removed. There is now only one entry, `xh-n-isisu`, in the lexicon.

Class `prov:Dictionary` is defined as: "an entity that provides a structure to some constituents, which are themselves entities. These constituents are said to be members of the dictionary" and the notion of 'dictionary' is extended further to include "a wide variety of concrete data structures, such as maps or associative arrays" (W3C, 2013a).

3.9 Generation of Linked Data

A digital repository is defined by IMS Global Learning Consortium (2003, cited in Simons & Richardson, 2013:2) as "any collection of resources that are accessible via a network without prior knowledge of the structure of the collection."

One or more lexicons (resources) with lexical entries (items in a resource) in a single namespace meets this definition, and if managed by an institution within the field of LIS, is expected to be compliant with the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), a framework developed for repository interoperability, as the first step when ensuring interoperability (Simons & Richardson, 2013:14 8-151). The metadata record of an item must be an XML document which "**must** use the UTF-8 representation of Unicode", expressed in Dublin Core format, with URIs as unique identifiers ("Open Archives Initiative …", 2015). The record can be served using the same use cases for URIs as previously defined, with a representation of the metadata record configured specifically for XML-aware clients ("Sample OAI-PMH requests", n.d.). Whilst further detail is outside the scope of this study, if OAI-PMH compliance is required, then the URIs and the metadata must be UTF-8 encoded.

Literature makes limited mention of generation and publication of RDF data (Vila-Suero et al., 2014:113; Ehrmann et al., 2014:406; Gracia, 2018:5-7), although Ehrmann et al.,

when discussing the ways BabelNet is served on the web, talk of downloadable RDF dump files (which at time of writing, no longer appear to be available). Gracia et al. (2018:6) talk of an RDF file per lexicon, loaded into a Virtuoso[20] triple store, and the RDF data accessible by way of a SPARQL endpoint and a Linked Data interface developed using Pubby[21]. In a 2017 presentation, Gracia explored the topic further, advising that the dataset should be registered in Datahub.io, the data stored in a SPARQL store, with "a mechanism to make our URIs dereferenceable" by means of a Linked Data front-end, or served as files using a common web server (Gracia, 2017).

The versioning requirements identified in the previous section suggest the following approach to publication:

- As a lexical entry (or its senses or translation relations of which one of its senses of is the source) changes, a new file in the various formats required should be generated. **U1** would always point to the latest version of the lexical entry. This can be an automated task, scheduled to run at a predetermined time.

- Due to the way provenance is recorded for lexicons using PROV-O, the lexicon file may need to be published more than once a day. The suggested URI pattern to account for this is:

  U3: `{http(s)://}{Base URI}/{Resource Path}/{Resource ID}/{Version ID}/{Version Number}`

  where `{Version ID}` is the date, and `{Version Number}` is an incremental number relative to the `{Version ID}`.

- Lexical entries are members of the lexicon collection, and any changes to members (by way of insertions or deletions), should generate a new version of the lexicon file, using the same principle described above. As a lexicon modelled using Ontolex-Lemon can only contain lexical entries of one language, if the dataset has two or

[20] https://virtuoso.openlinksw.com/
[21] http://wifo5-03.informatik.uni-mannheim.de/pubby/

more languages, then there would be two or more lexicons and the process would be repeated per lexicon.

- The files representing the latest version of the lexicon and its lexical entries can be concatenated and compressed to create a data dump.

These files would be static RDF files stored on the web server, identified by Heath and Bizer as "the simplest way to publish Linked Data", and the file generation can be automated (Heath & Bizer, 2011:72). For a SPARQL endpoint, the data dumps can be manually uploaded to Dydra[22], a cloud-hosted RDF platform, using their online form ("Dydra", 2011; Dydra, n.d.).

3.10 Summary

This chapter described the processes associated with LLD generation: from URI design, to modelling a lexicon, a lexical entry, its senses, and translation relations, as well as the associated metadata; to versioning; and publication and generation of the datasets. Although a bilingual resource was the primary consideration, the approach for multilingualism was also considered.

The research questions, Q1 and Q2, were also addressed:

Q1: *"How does one construct the LLDF so as to allow for extensibility from a bilingual to a multilingual resource?"*

This question was primarily addressed with the URI design use cases, demonstrating the extensibility of URI patterns to additional languages. The modelling of lexicons and lexical entries were also shown to be extensible to a multilingual environment, with the relationships between the lexical senses of three languages considered. The *vartrans* module of Ontolex-Lemon, designed to model translation relations, was discussed in detail.

[22] https://dydra.com

Q2: *"How does one construct the LLDF to allow for change, tracking provenance of each change?"*

This was primarily addressed by demonstrating the modelling of provenance for lexicons, lexical entries, their senses, and associated translation relations. Generation of the RDF data for an evolving dataset was also considered.

The following chapter analyses the selected case study, using Ontolex-Lemon, the selected ontologies and vocabularies identified in Chapter 2, and aspects of the methodological guidelines described in this chapter.

Chapter 4 Analysis using Ontolex-Lemon

4.1 Introduction

As described in Section 1.2.1, EXDN is a dictionary in printed form. Although the digitisation process of a lexicographic resource is outside the scope of this study, Figure 4-1 illustrates the workflow conducted to digitise EXDN (legend in Appendix C).

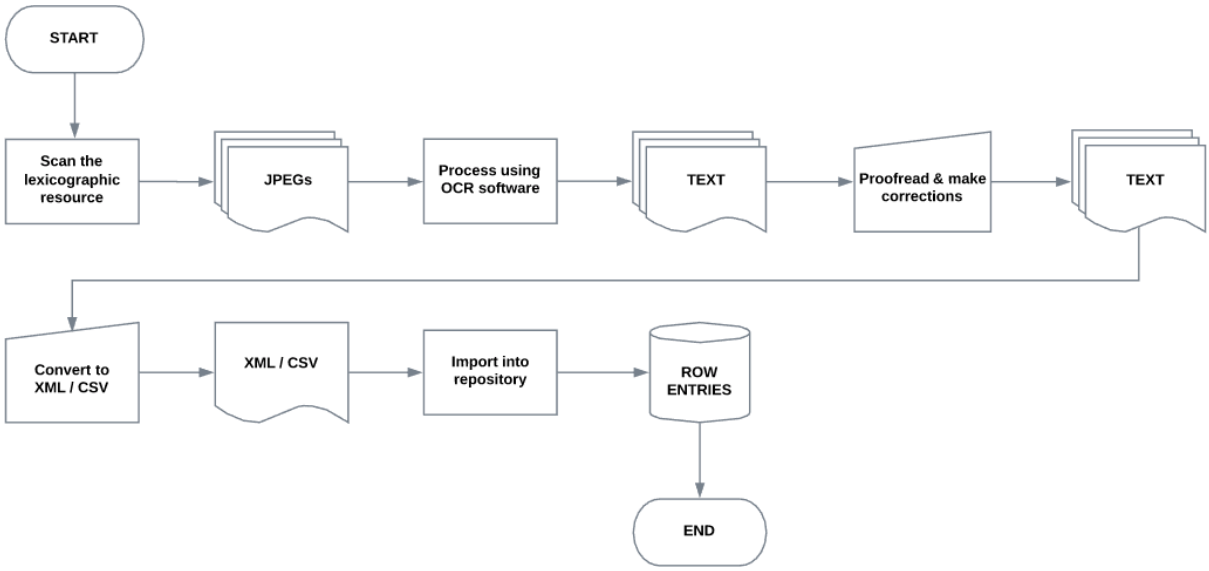


Figure 4-1: Workflow illustrating the digitisation process - from a dictionary to row entries in a database

A DWS was custom-developed using PHP and a MySQL relational database; the purpose of which was to manage the lexical entries, to prepare the lexical entries for publication, and to generate the RDF in the formats required for the published lexical entries. Figure 4-2 illustrates the publishing process of a single lexical entry.

Before a lexical entry can be published, the language needs to be identified, the lexical entry type needs to be determined, the POS identified, a resource identifier created, a minimum of one sense created, and a lexical concept if not previously created. The type

can be an affix, word, or multiword expression and corresponds to the subclass of Class Lexical Entry in Ontolex-Lemon, or it can be an acronym (`lexinfo:Acronym`) or stem (`mmoon:Stem`). The POS corresponds to the LexInfo vocabulary of parts of speech. The resource identifier takes the form identified in Section 3.2.4, and once the requirements for publishing a lexical entry have been fulfilled, it is automatically scheduled for inclusion when the next version of the lexicon is generated.

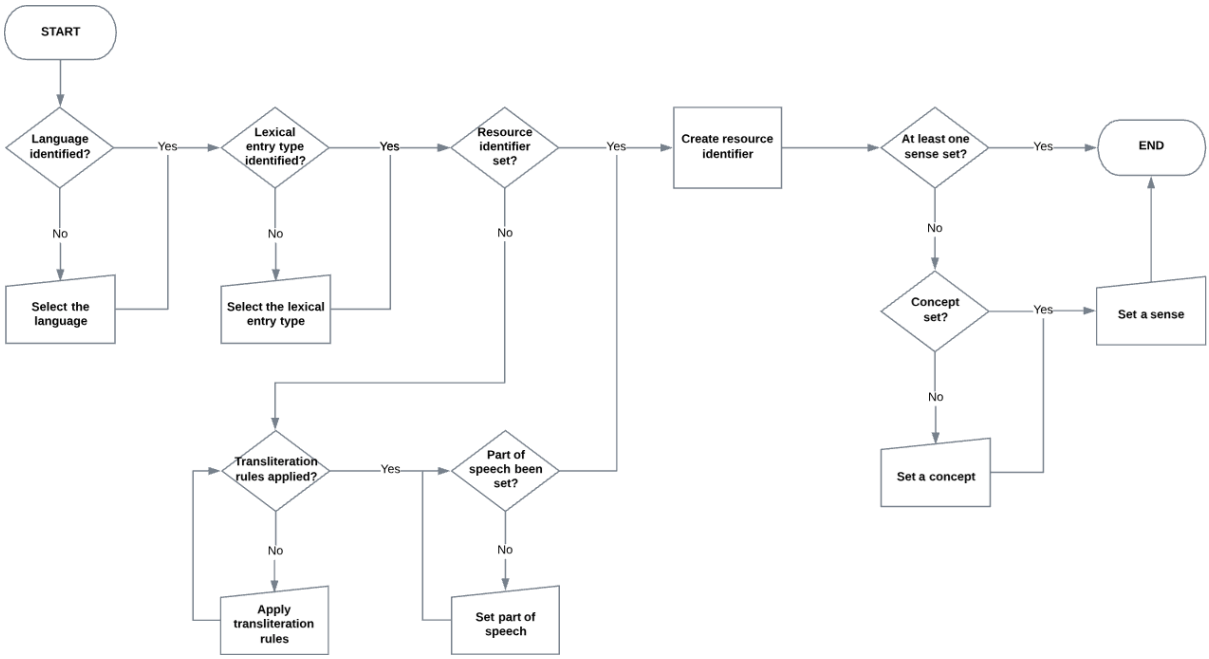


Figure 4-2: Workflow illustrating the publishing process

This process can be applied to any lexicographic resource, although depending on the age of the resource, transliteration rules may need to be defined per language. For English, the source language, the following transliteration rules were applied:

TR1: æ → ae

TR2: œ → oe

For isiXhosa, the target language, the following transliteration rules were applied (as per Doke, 1954:91):

TR3: ɓ → b (implosive)

TR4: ʃ → sh (voiceless prepalatal fricative)

TR5: ɽ → r (voiceless velar fricative)

The following changes were also implemented for each lemma (either programmatically or by manual input):

- Conversion to lowercase.
- "." was removed at the end of the lemma.
- Plural forms converted to singular.
- Synonyms were removed and created as separate lexical entries, with a relation to the original lexical entry indicated.

Figure 4-3 shows a screenshot of a published lexical entry, *abdomen*, within the DWS.

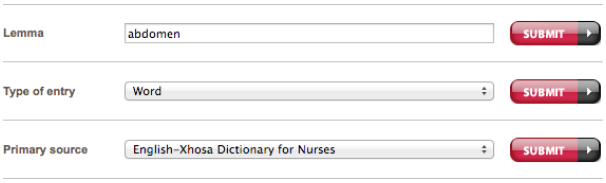


Figure 4-3: A lexical entry in the DWS

4.2 Identification of the use cases

Three approaches can be taken to convert data to RDF (Hyland & Villazón Terrazas, 2011):

1. Automatic conversion, sometimes called *triplification*
2. Partial scripted conversion
3. Modeling by human and subject matter experts, followed by scripted conversion.

For EXDN, Approach 3 was taken, using the modelling examples defined in Chapter 3.

Each lexical entry (excluding the lemma sign) was translated using Google's Cloud Translation API[23], which has two models: Phrase-Based Machine Translation and Neural Machine Translation (Google Cloud, 2018). Each lexical entry was translated using both models. Both models rendered results which were either incorrect or not of a sufficient standard to present to the end-user, although in some instances the translations provided enough information to aid sense disambiguation. Figure 4-4 shows the translation of *Abdomen*.

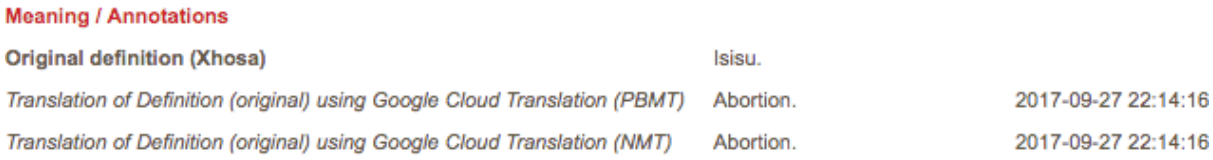


Figure 4-4: Translation using Google's Cloud Translation API

With sense disambiguation, the following forms of each article were identified (Gouws & Prinsloo, 2005:93-94,125):

F1: The article offers a restricted treatment of the lemma sign.

In a monolingual dictionary, this form is often employed for a lesser-known synonym and is intended to serve as a cross-reference entry (Gouws & Prinsloo, 2005:94). For a bilingual dictionary, "where a source language item, represented by the lemma sign, is co-ordinated with a single target language item", this is a translation equivalent, with a full equivalence relation on both the lexical and the semantic levels (Gouws & Prinsloo, 2005:154).

An example of this form is the article "Abdomen" (1935:1):

> **Abdomen.** Isisu.

F2: The article offers a paraphrase of meaning of the lemma sign.

[23] https://cloud.google.com/translate/

A paraphrase of meaning is also referred to as a lexicographic definition (Gouws & Prinsloo, 2005:143). However, in a bilingual dictionary, an article typically includes a collection of translation equivalents, for which there can be full or partial equivalence for each (Gouws & Prinsloo, 2005:151-155). If only a lexicographic definition is present in the article of EXDN, it is presumed (until clarified further) there is zero equivalence, and the definition serves to bridge the lexical gap between the source and target language (Gouws & Prinsloo, 2005:158).

An example of this form is the article "Fumigation" (1935:35):

> **Fumigation.** Ukuqhumisa egumbini nge-*sulphur* nezinye izinto.

F3: The article contains a cross-reference entry.

An example of this form is the article "Change of life" (1935:18):

> **Change of life.** Khangela *Menopause*.

F4: The article offers a comment on semantics.

An example of this form is the article "Aqua or Aq." (1935:7):

> **Aqua or Aq.** (*Latin for water*). Amanzi.

From this and the data from EXDN, the following use cases have been identified for modelling:

M1-4: Modelling the forms **F1** to **F4**,
M5: Modelling a plural form for an African language,
M6: Modelling an outdated lexical entry, and
M7: Modelling synonymy and relatedness.

Before modelling can begin for the identified use cases, the lemmatisation approach should be considered.

4.3 The lemmatisation approach

Lemmatisation is defined as a selection of a form, be it a word, a multiword expression, an affix or a stem, as the starting point for retrieval of information for that article (Gouws & Prinsloo, 2005:67). The lemmatisation approach is the approach considered for the central list of the dictionary, and according to Zgusta, all cases of the same type should be treated in the same way, unless specific reasons are given (1971:314).

In this study, the lemmatisation approach is chiefly focussed on Bantu languages, as languages like English usually follow the word tradition. Bantu languages are characterised by a noun class system with concordial agreement, and for verbs, numerous derivations of the verb stem can exist, by way of prefixes and suffixes (Gouws & Prinsloo, 2005:68-69; Chavula & Keet, 2014:2).

When compiling a dictionary, the lexicographer has to consider the words "which are most likely to be consulted by the target user and to lemmatise them in a user-friendly way" (Gouws & Prinsloo, 2005:39). When considering the discussion by De Schryver of target users expecting derived forms (2010:162), the "target user" is important here. As demonstrated by way of example for the stem "-su", in a bilingual dictionary, for example, isiXhosa→English, a mother tongue speaker may expect to look up the stem "-su", and an inexperienced (mother tongue or foreign) user of the language may expect to look up one of its derived forms, such as "isisu". Although the size of a physical dictionary constrains the lemmatisation list, support for lemmatising highly used derivations is given by both Zgusta (1989:300, cited in Gouws & Prinsloo, 2005:74), and Gouws and Prinsloo (2005:68-74), unlike that of Doke, who deemed dictionaries which took the word approach as "large vocabularies" (1954:18). Ten years after Doke, Benson followed up with a "cardinal principle" (1964:82, cited in De Schryver, 2010:162): "everything which needs to be said about a stem or root should be channelled into one single full article."

Some twenty years later however, Benson acknowledged that "the user must be able to identify the stem, which given sometimes complex morphophonemics of Bantu languages may not be easy" (1986:3-4, cited in De Schryver, 2010:163).

For the bidirectional *English-Xhosa / Xhosa-English Dictionary* published by Pharos Dictionaries, the second edition published in 1998, and on its sixteenth printing by 2014, the stem approach for nouns and verbs was adopted in its Xhosa-English section, although the front matter, when describing the lemmatisation approach, ended with: "to make sure of using words correctly it is often wise to compare a Xhosa word found in the English-Xhosa section with the meaning given in the Xhosa-English section (if it is to be found there)" (English-Xhosa / Xhosa-English Dictionary, 2014).

The *Eiwanika ly'Olusoga*, a Lusoga (a Ugandan Bantu language) dictionary compiled in 2008 by Nabirye, took a word lemmatisation approach (2009, cited in De Schryver, 2010:164), and in 2010, the *Oxford Bilingual School Dictionary: Zulu and English* took a word-based approach for nouns (De Schryver, 2010:164). Prinsloo (2010:760-761), declared the latter dictionary "excellent" and went on to say that "no lexicographer or lexicographic unit to my knowledge [has] thus far dared to break this almost sacred tradition of stem lemmatisation for nouns."

He commented on the likely argument of overcrowding articles if the full forms of nouns were lemmatised, and concluded that users were "unlikely to find it disturbing in any way" (Prinsloo, 2010:760-761). Unlike printed dictionaries, web lexicographic resources are not limited by physical space, and a hybrid approach can be considered: the lemmatisation of both stems and frequently used derived forms (based on frequency of occurrence in a selected corpus). If derived noun forms are lemmatised, then the lemmatisation of both singular and plural noun forms will also need to be considered (Gouws & Prinsloo, 2005:76). For English, plural noun forms are typically *noun*+s (Jurafsky & Martin, 2009:82), however, in an African language, the prefix changes, rendering a different lemma, which (a) may provide a more frequently used form, or (b) which an inexperienced language user may use when looking up a word (Gouws & Prinsloo, 2005:77-78). For lemmatisation of verbs, according to Gouws and Prinsloo (2005:79-80), "a huge number of prefixes, up to more than *4,000 per verb*, combine freely and productively with verbs" (researcher's emphasis), and they therefore recommend the stem tradition for verbs. For inexperienced language users, isolating the stem, particularly for words with a conjunctive orthography, can be problematic (Gouws &

Prinsloo, 2005:80). For the lemmatisation of nouns, Gouws and Prinsloo recommend the word approach, lemmatising both the singular and plural forms (2005:84).

EXDN is a unidirectional dictionary, but for articles with full equivalence, such as *abdomen→isisu*, it is assumed that if directionality were reversed, it too would have made use of the word lemmatisation approach. Using the stem tradition does accurately represent the language; however, for the word tradition, a user would be shown a translation equivalent that, according to Zgusta (1971:32) "is always a possible and sometimes the best possible choice for insertion into a real sentence."

The following additional use cases for modelling have been identified:

M8: Using a stem as the lemma,
M9: Using a derived noun as the lemma,
M10: Modelling a lexical entry for a derived noun, with the plural form as the lemma,
M11: Modelling a translation relation between a source and target sense, which do not share the same lemmatisation approach.

4.4 Modelling the article: *Breath* (1935:18)

For the modelling that follows, only triples relevant to modelling the use cases are shown.

The points to be modelled:

- It has a translation equivalent: *umphefumlo*
- *umphefumlo* is a derived noun
- The stem is: -phefumlo
- The plural of *breath* is *breaths*
- The plural of *umphefumlo* (NC 7) is *imiphefumlo* (NC 8) (see Appendix D for a description of the noun classes, abbreviated here as NC)

The following use cases are addressed with the modelling of this article:

M1: The article offers a restricted treatment of the lemma sign

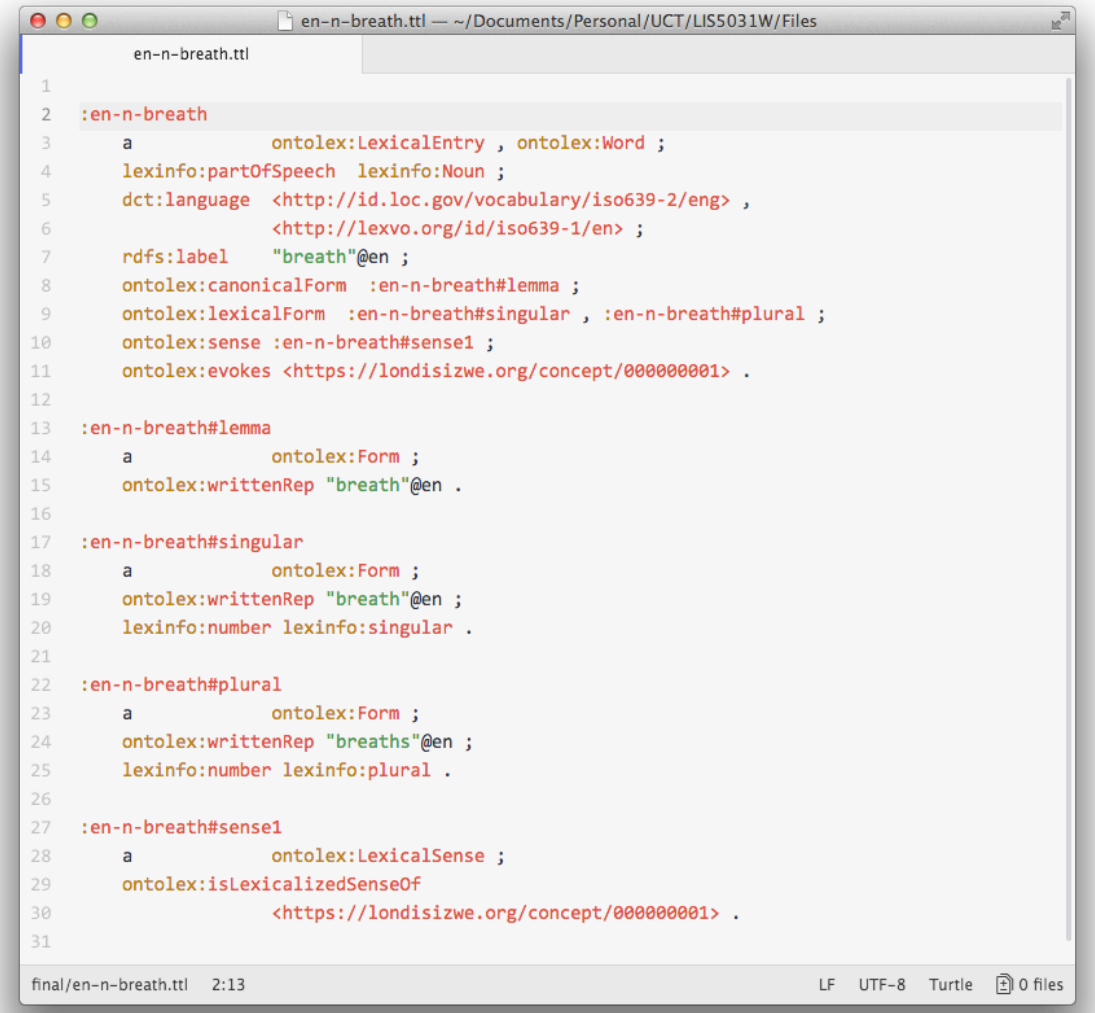
M5: Modelling a plural form for an African language

M8: Modelling a lexical entry with a stem as the lemma

M9: Modelling a lexical entry with a derived noun as the lemma

M11: Modelling a translation relation between a source and target sense that do not share the same lemmatisation approach.

4.4.1 Describing the lexical entry: `en-n-breath`

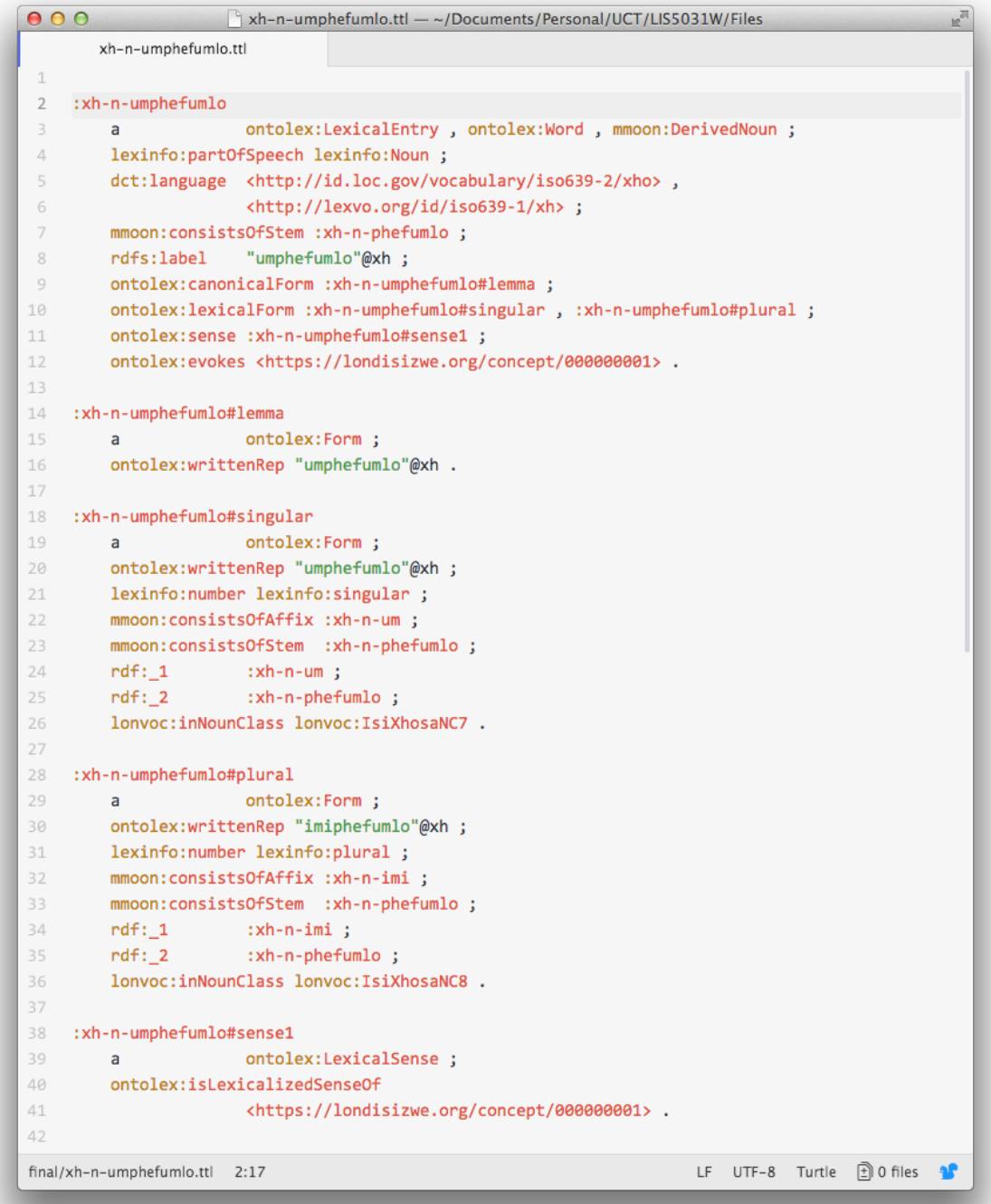


Code Example 4-1: Describing the lexical entry en-n-breath

Where:

- Lines 17-20: describe the singular form
- Lines 22-25: describe the plural form
- Line 29-30: links a lexical concept to this sense (inverseOf)

4.4.2 Describing the lexical entry: `xh-n-umphefumlo`

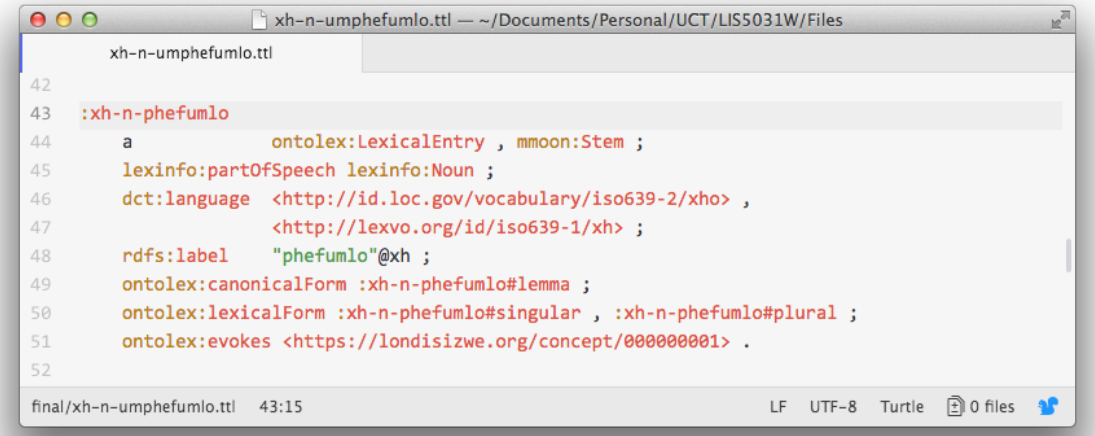


Code Example 4-2: Describing the lexical entry xh-n-umphefumlo

Where:

- Line 3: indicates that it is a derived noun
- Line 7: identifies the lexical entry of the stem
- Lines 18-26: identifies the singular form of this lexical entry
- Lines 22-23: indicates the affix and the stem of this form
- Lines 24-25: shows the order in which the derived noun is composed
- Line 26: indicates the noun class to which this form belongs
- Line 28-36: identifies the plural form
- Lines 32-33: indicates the affix and the stem of the plural form
- Line 36: indicates the noun class of the plural form.

4.4.3 Describing the lexical entry: `xh-n-phefumlo`

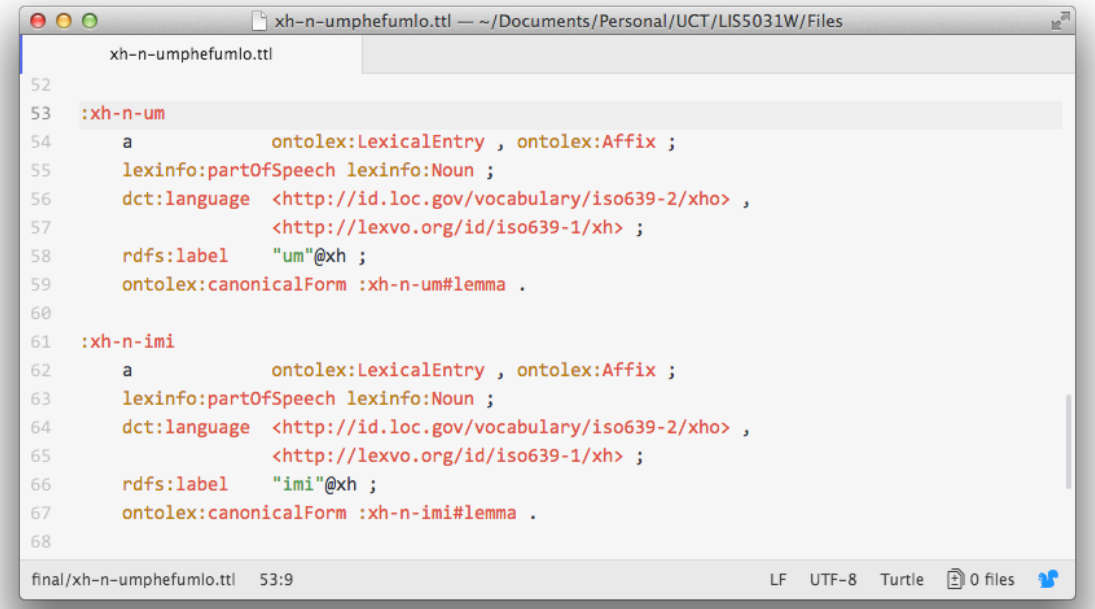


Code Example 4-3: Describing the lexical entry xh-n-phefumlo

Where:

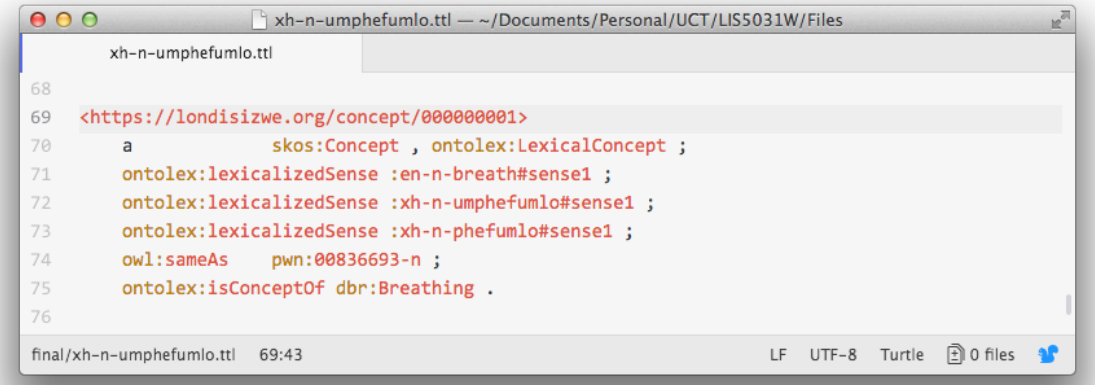
- Line 44: identifies this lexical entry as a stem

4.4.4 Describing the prefixes: `xh-n-um`, `xh-n-imi`



Code Example 4-4: Describing the prefixes xh-n-um, xh-n-imi

4.4.5 Describing the concept: `000000001`



Code Example 4-5: Describing the concept 000000001

Where:

- Lines 71-73: indicates the lexicalised senses of this concept (`en-n-breath#sense1`, `xh-n-umphefumlo#sense1` and `xh-n-phefumlo#sense1`).

- Line 74: identifies this concept as the same as the PWN sense
- Line 75: identifies the subject as a concept of a DBpedia resource

For the use case **M1**, when applied to a bilingual resource, translation equivalence is easily modelled between the senses of two separate lexical entries using the *vartrans* module. However, as there is not always an ontology entity for denoting the meaning of a lexical entry, and when there are entities available, they are not granular enough, the researcher felt it better to define a `ontolex:LexicalConcept`, which is a subclass of `skos:Concept`, and identify the concept as the `ontolex:lexicalizedSense` of the appropriate sense; and within the `ontolex:LexicalConcept`, it is indicated to be a concept of an ontology entity (Cimiano, McCrae & Buitelaar, 2016), thus:

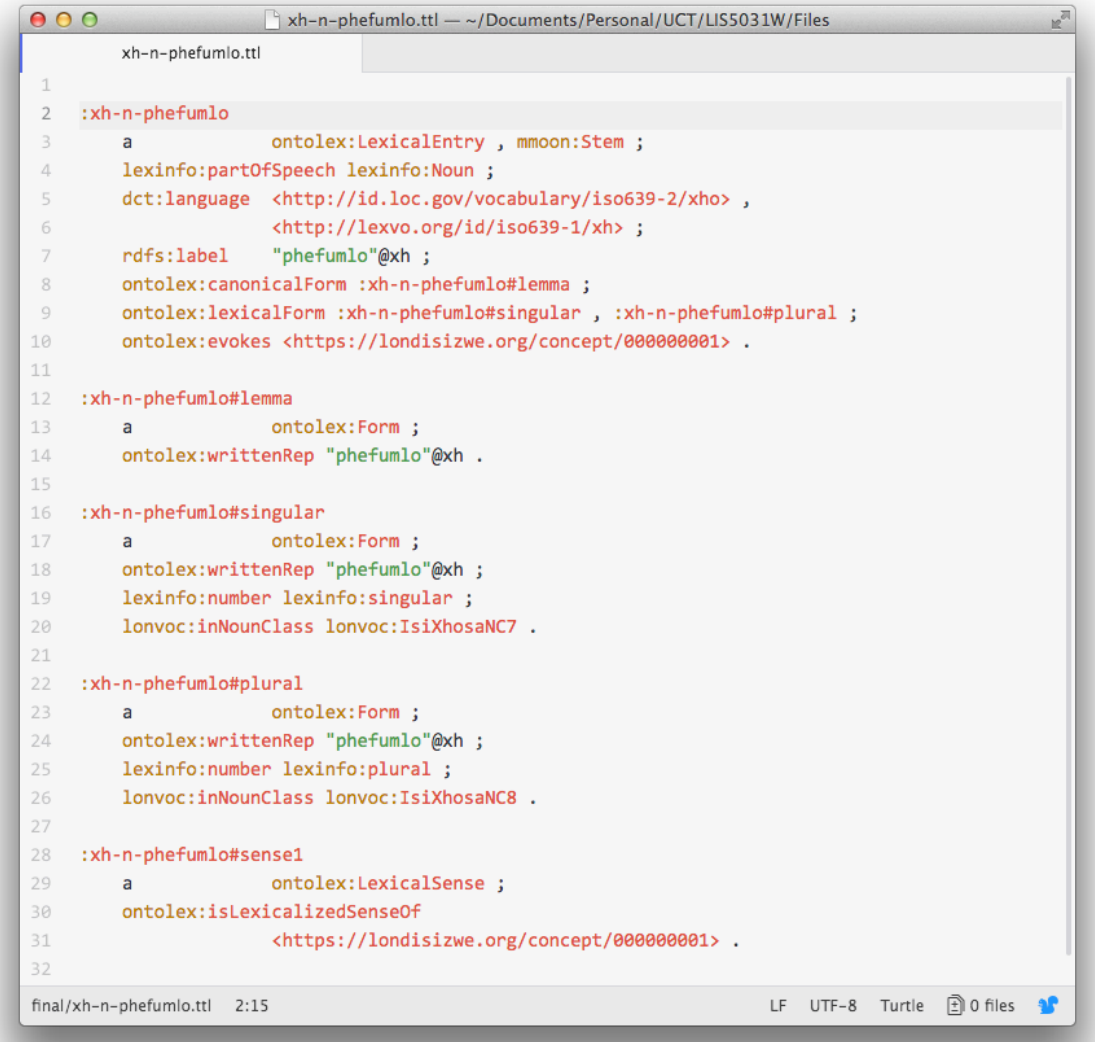
- it may not be necessary to declare translation relations from a source to a target language, which can be onerous for a multilingual resource;
- any sense from any language which shares the same `skos:Concept`, is thus deemed equivalent.

Using `skos:Concept` as a reference to the sense is supported in the literature, with use by Bosque-Gil et al. when applying Ontolex-Lemon to Terminesp (2015:290), and Bosque-Gil, Gracia and Gómez-Pérez (2016:20). The advantages of creating a concept are:

- lexicographic definitions (**F2**), cross-reference entries (**F3**), comments on semantics (**F4**), and usage examples can all be described within the concept, and not only to a particular sense;
- this separates the concept from the sense, thus allowing for easy reuse by external resources, and adhering to the *semantics by reference* principle (Bosque-Gil, Gracia & Gómez-Pérez, 2016:20);
- using machine translation or manual translation methods, one is able to translate the lexicographic definitions, usage examples, and comments on semantics, so even if there is a lexical gap for a target language, or the equivalent sense has not yet been described within the RDF, the meaning of a lexical entry in a source language could still be understood.

The lexical concept resembles a WordNet *synset*, although instead of modelling sets of similar terms (synonyms), the lexical concept applied here models sets of equivalent senses across languages (Bosque-Gil et al., 2015:290).

For use case **M5**, the plural form of an African language is modelled from lines 28-36. Because the noun class changes for the plural form, a vocabulary has been created which defines the `lonvoc:inNounClass` property, as well as classes for each noun class of the isiXhosa language (see Appendix E). This vocabulary can be extended to other African languages as well.



Code Example 4-6: Modelling the lexical entry xh-n-phefumlo

For use case **M9**, it is identified as a derived noun using the `mmoon:DerivedNoun` class (line 3). The stem of the derived noun is indicated using the `mmoon:consistsOfStem` property. Because of the change of prefix between a singular and plural form, the affixes are defined within the singular and plural form for the lexical entry, with `rdf:_1`, `rdf:_2` used to indicate the order.

For use case **M8**, modelling the stem as the lemma, assuming there are no derived nouns, the lexical entry `xh-n-phefumlo`, can be modelled as shown in Code Example 4-6 (its derivation from the verb -phefumla is not modelled here).

For use case **M11**, namely modelling a translation relation between senses which do not share the same lemmatisation approach: because a sense is identified as a `ontolex:isLexicalizedSenseOf` a concept, any other sense, be it a derived word or a stem, which are lexicalised to the same concept are equivalents.

The MMoOn ontology has been used extensively by the researcher to model the isiXhosa lexical entries. McCrae and Gracia (2017), when discussing the future direction of Ontolex-Lemon, referred to a new module for morphology, based on the MMoOn ontology, so although a stem cannot be explicitly modelled using a subclass of `ontolex:LexicalEntry` in Ontolex-Lemon, this may be a possibility in the future.

4.5 Modelling the article: *Fumigation* (1935:35)

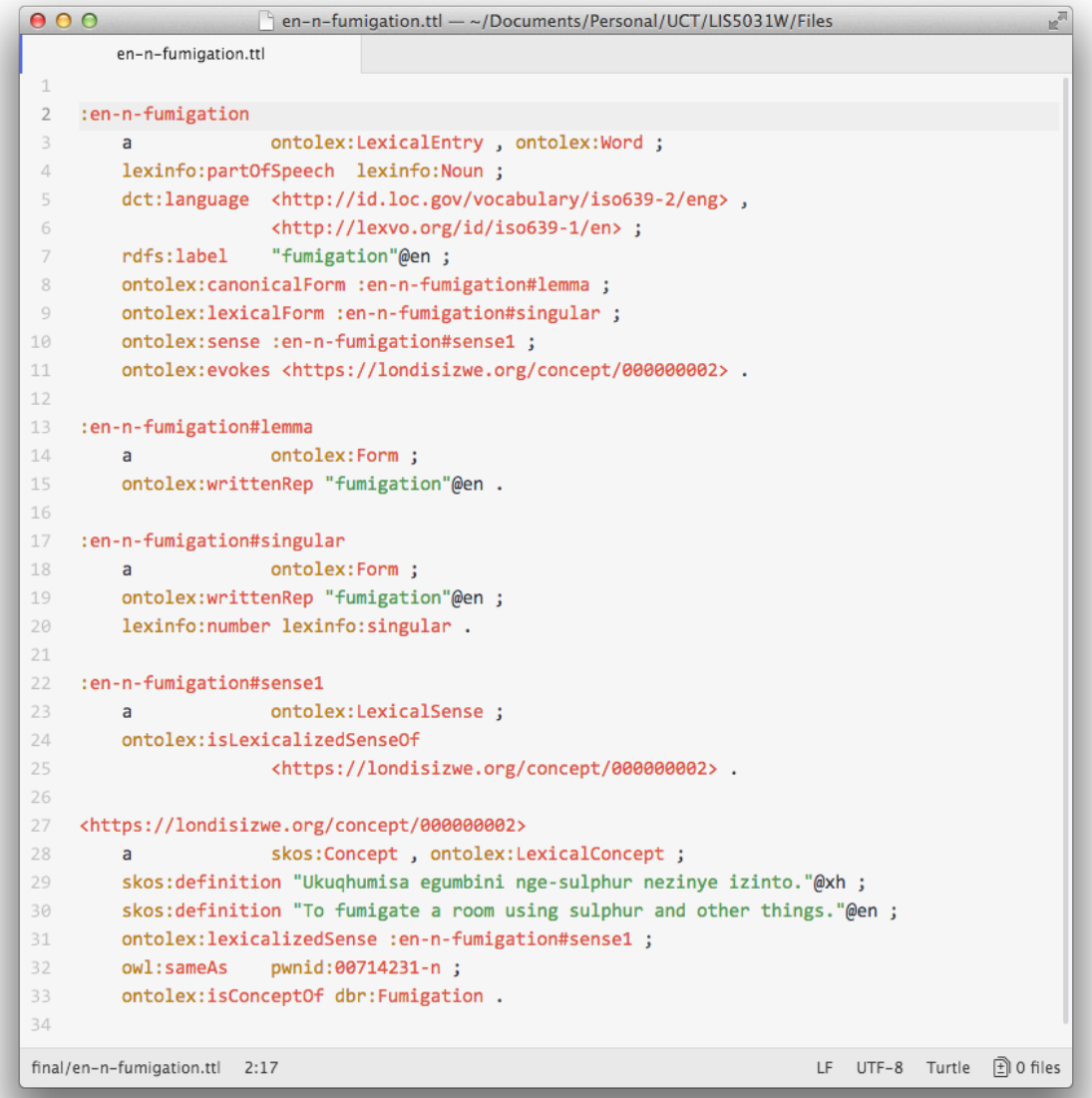
The points to be modelled:

- The lexicographic definition
- There is no translation equivalent

The following use case is addressed with the modelling of this article:

M2: The article offers a paraphrase of meaning of the lemma sign

4.5.1 Describing the lexical entry: `en-n-fumigation`



Code Example 4-7: Modelling the lexical entry en-n-fumigation

Where:

- This lexical entry only contained a lexicographic definition, so the isiXhosa translation equivalent cannot be modelled
- Line 29-30: indicates the definition for the lexical sense of the lexical entry, thus modelling **M2**

4.6 Modelling the article: *Change of life* (1935:18)

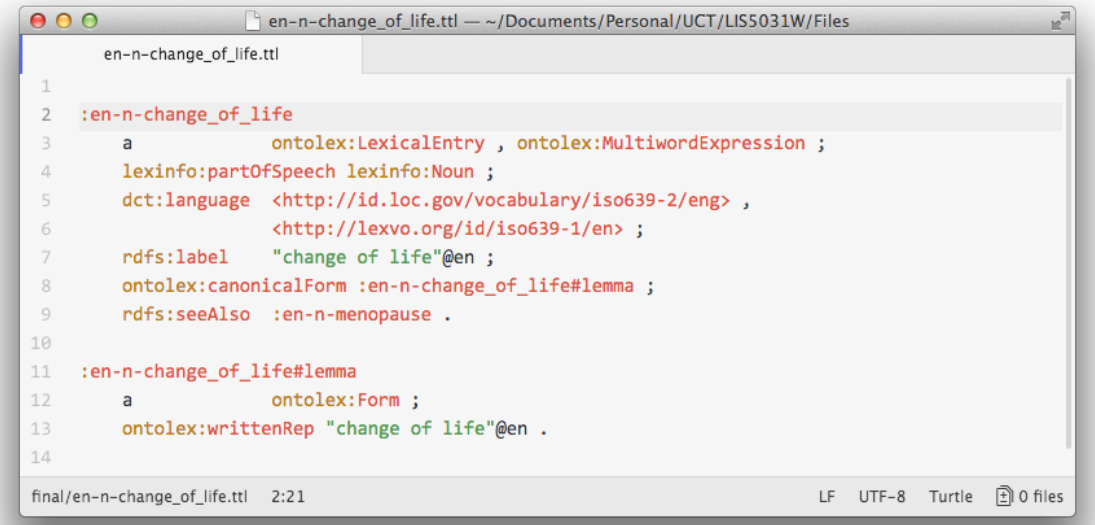
The point to be modelled:

- There is a cross-reference entry to *Menopause*

The following use case is addressed with the modelling of this article:

M3: The article contains a cross-reference entry.

4.6.1 Describing the lexical entry: `en-n-change_of_life`



Code Example 4-8: Describing the lexical entry en-n-change_of_life

Where:

- Line 9: indicates the cross-reference entry to `:en-n-menopause`, thus modelling **M3**

Because there are no senses modelled for this lexical entry, where the purpose is solely to indicate the cross-reference entry, `rdfs:seeAlso` is modelled at lexical entry level. However, if there were senses, and a cross-reference entry was only applicable to one of those senses, then it would be modelled within the sense.

4.7 Modelling the lexical entry: *Amanzi* ("Aqua or Aq.", 1935:7)

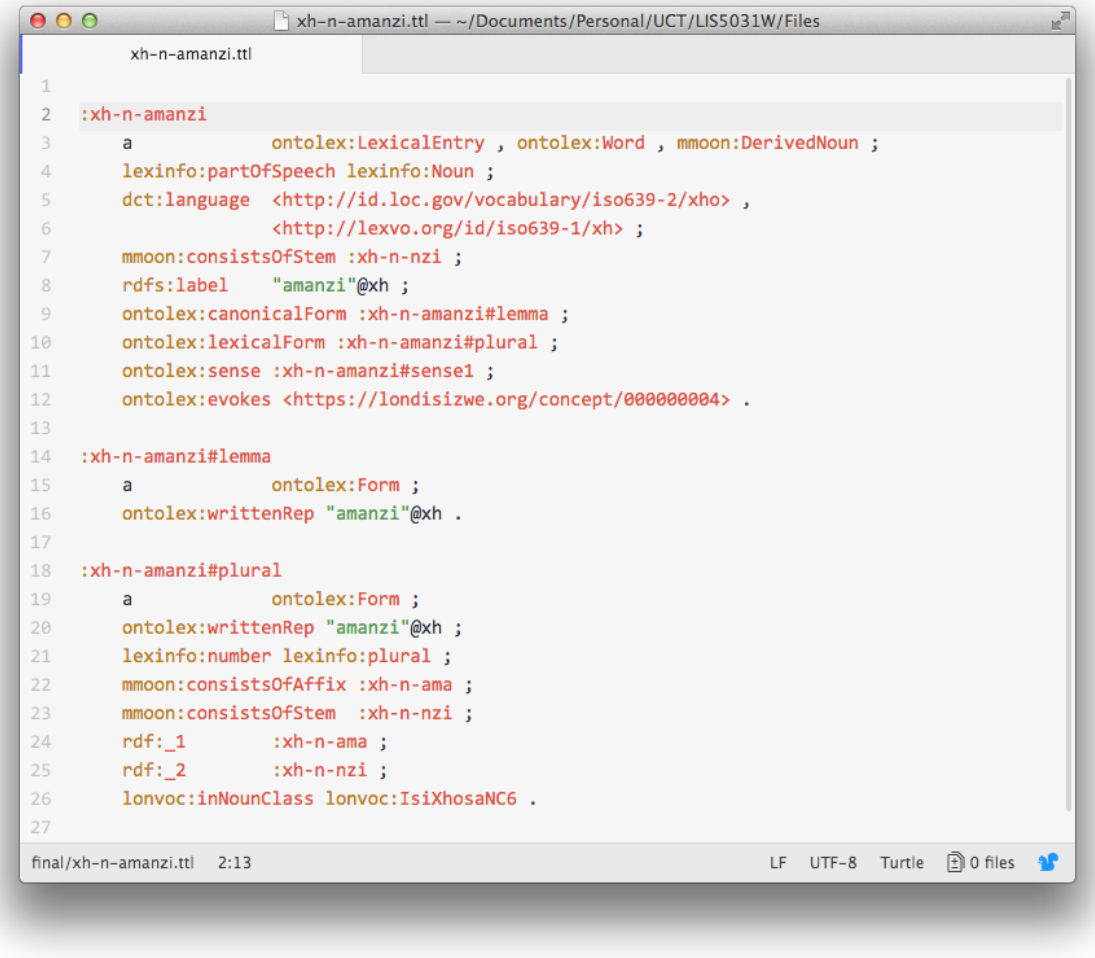
The points to be modelled:

- This is a derived noun from "-nzi"
- The plural form is the lemma (the noun being an invariant plural)

The following use case is addressed with the modelling of this article:

M10: Modelling a lexical entry for a derived noun, with the plural form as the lemma

4.7.1 Describing the lexical entry: `xh-n-amanzi`



Code Example 4-9: Describing the lexical entry xh-n-amanzi

Where:

- The plural form is the same as the lemma

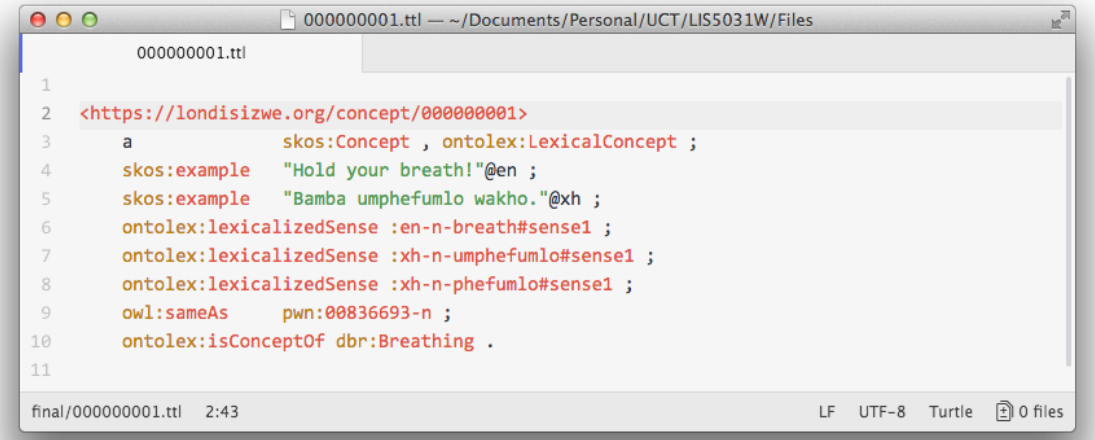
4.8 Modelling comments, definitions, scope notes, usage and examples

M4, a comment on semantics, and other forms of annotation, can be modelled using the following properties within the lexical concept (W3C, 2009):

- `rdfs:comment`
- `skos:definition`
- `skos:example`
- `skos:scopeNote`

`skos:example` is modelled below.

4.8.1 Describing the concept: `000000001`



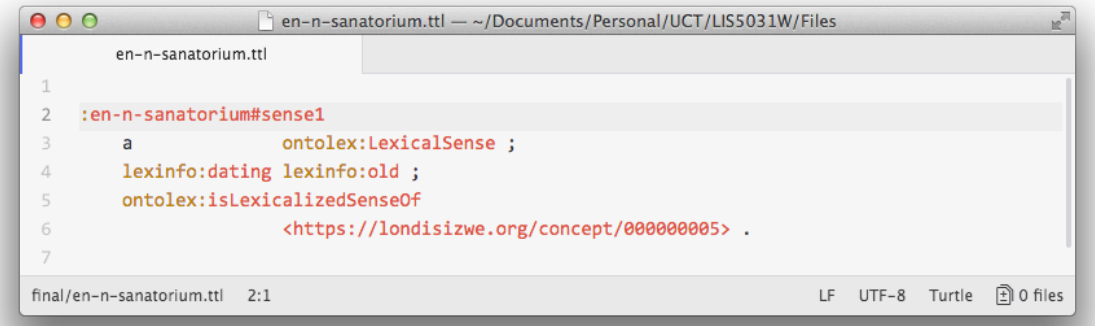
Code Example 4-10: Describing the concept 000000001

By including comments, scope notes, usage notes, and examples, context and cotexts can also be defined.

4.9 Modelling other forms

Modelling a lexical entry as outdated (**M6**) is done at sense level:

4.9.1 Describing the sense: `en-n-sanatorium#sense1`



Code Example 4-11: Describing the sense en-n-sanatorium#sense1

Where:

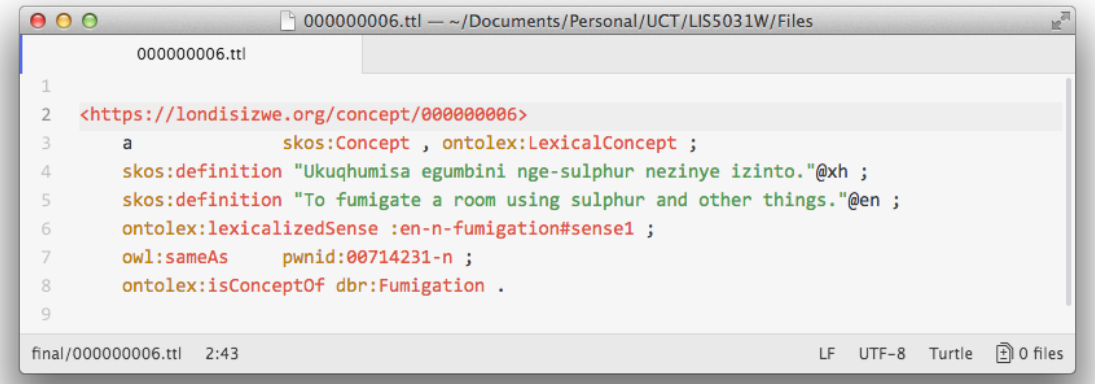
- Line 4: indicates the outdated nature of the sense

Synonymy and relatedness (**M7**) can be modelled using SKOS within the lexical concept: `skos:broader` (hypernymy), `skos:narrower` (hyponymy) and `skos:related` (synonymy).

4.10 Modelling annotations

As shown in the previous examples, resources can be annotated using `rdfs:comment` and `rdfs:label` (Heath & Bizer, 2011:59). However, when revisiting the lexical concept for `en-n-fumigation`, another form of annotation is also necessary.

4.10.1 Describing the concept: `000000006`



Code Example 4-12: Describing the concept 000000006

Where:

- Line 4: has the English word *sulphur* in the isiXhosa definition

The implication of this is as follows:

- The definition should be annotated by linking the term to an entity, be it an ontology entity, a lexical entry defined for *sulphur*, or a lexical concept. Annotations can be added using the W3C Web Annotation Vocabulary[24] or the NLP Interchange Format (NIF)[25], or any other annotation model (Cimiano, McCrae & Buitelaar, 2016).
- Defining translation relations between senses becomes necessary, instead of relying solely on equivalence due to a shared lexical concept, because a machine translation, may be inaccurate if the source language is not identified.

Returning to the four levels of data defined in Section 3.8, the Primary Level can be updated to include lexical concepts:

Primary level: RDF triples representing the lexicons, lexical entries and lexical concepts

[24] https://www.w3.org/TR/annotation-vocab/
[25] http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#

Figure 3-3 has also been updated accordingly, as shown in Figure 4-5.

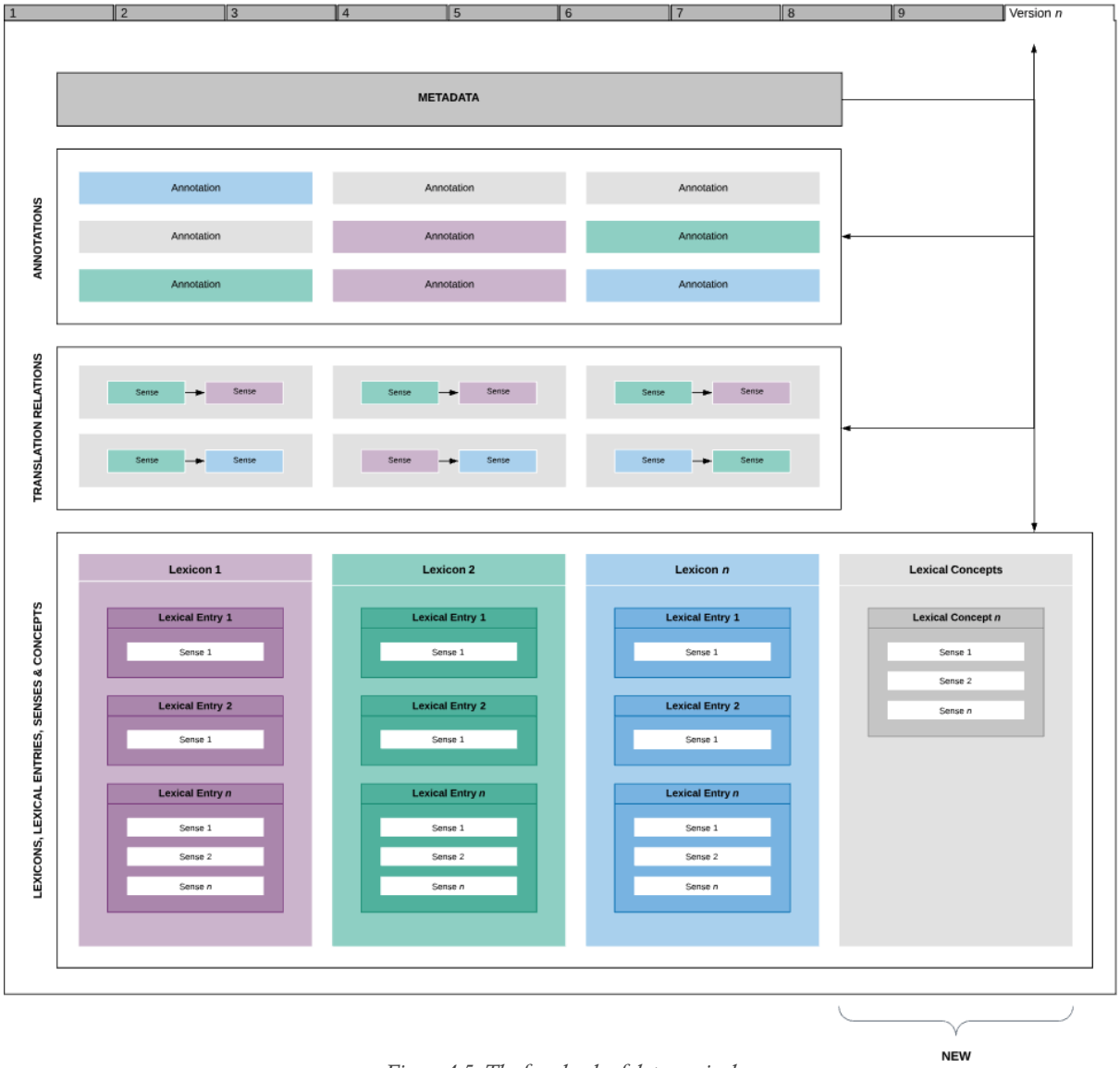


Figure 4-5: The four levels of data, revised

4.11 Summary

Whereas a dictionary is typically hierarchical, the RDF data model is a network or graph structure (Lowe & Hall, 1999:62-64). As a result, when converting a dictionary, reinterpretation of the data is required to convert it from a tree-view to a graph-view, and if the representation of information exactly as in the original resource is required, then

RDF may not best suit the task. An example of reinterpretation is the dictionary article, "**Change of life.** Khangela *menopause.*" (1935:18), which is represented as:

Subject	*Predicate*	*Object*
`:en-n-change_of_life`	`rdfs:seeAlso`	`:en-n-menopause`

The representation of hierarchy in RDF, be it in the form of sub-senses or inflection with multiple affixes attached to a word stem, has been identified as a modelling challenge by Gracia, Kernerman and Bosque-Gil, as has the modelling of translations and examples (2017:5). The lemmatisation approach and annotations within a multilingual environment are another modelling challenge identified in the case study. A lexicography module for Ontolex-Lemon is in progress, which will hopefully address the issues identified (Bosque-Gil, 2017). Although Ontolex-Lemon takes a modular approach, as its range of data representation extends, provenance and versioning, discussed in Chapter 3, will be necessary to accommodate any change to the RDF data representation.

The research question, Q3, was addressed in this chapter:

Q3: *"How does one construct the LLDF for translation equivalents, which may have a differing lemmatisation approach for nouns and verbs?"*

This question was primarily addressed with a discussion of the lemmatisation approaches for African languages. The modelling of derived nouns, stems and affixes were demonstrated using data from the case study, with senses of words, derivations and their stems linking to a lexical concept that "represents *a mental abstraction, concept or unit of thought that can be lexicalized by a given collection of senses*" (Cimiano, McCrae & Buitelaar, 2016).

The final chapter continues with the theme of multilingualism and ends with an overview of the construction of an LLDF for bilingual lexicographic resources, compiled from the preceding chapters; a discussion of the findings; conclusions, and future work.

Chapter 5 **Discussion & Conclusions**

5.1 Introduction

Franceschini (2009:33-34, cited in Aronin & Singleton, 2012:6-7) offered an overarching definition of multilingualism as

> the capacity of societies, institutions, groups and individuals to engage on a regular basis in space and time with more than one language in everyday life.
>
> Multilingualism is a product of the fundamental human ability to communicate in a number of languages. Operational distinctions may then be drawn between social, institutional, discursive and individual multilingualism.

Within the domain of digital libraries, multilinguality has been defined as "multilingual information access" (Diekema & Eccles, 2012:166). Within HLT, Kay defines multilinguality as "a characteristic of tasks that involve the use of more than one natural language" where it is "more than just the preparation of parallel texts" (1997). In this study, multilinguality shares the digital libraries' definition, namely multilingual information access. However, due to the rare full equivalence between languages, this accessibility across two or more languages may be partial only.

McArthur described lexicography as the containerisation of knowledge (1986:130), with dictionaries, typically in print form, as the containers (Gouws, 2018:236; Nkomo, 2010:373). In a digital context, the language(s) contained within a single printed dictionary, bounded by its medium, no longer need to remain as "isolated entities" (Schoonheim, 2014:2). However, digitisation does not ensure multilinguality and unless expressly coded for, the digitised data will remain bound in its silo as "'monolingual islands' of data that do not interoperate" (Ehrmann et al., 2014:402). Using RDF, a lexicalised sense of any language can link to any lexical concept (which are essentially language-neutral) for which there is a shared conceptualisation, thus enabling lexicalised senses from other languages, as well as their near-synonyms, to be accessible.

This chapter will focus on the modelling of languages with click consonants not represented by the Roman alphabet, thereby addressing the research question Q4. To

conclude, an overview of the construction of the LLDF discussed in the preceding chapters will be presented: thereby answering Q0, with a critical evaluation of the model for Q1 – Q4, as well as the consideration of future work. In closing, the model is considered through a lens of 'context'.

5.2 Modelling click languages

Doke defined a click as "an injected consonant produced by a rarefaction between two points of closure", with one point of closure being the tongue (1954:35). IsiXhosa has three clicks, each represented by the Roman alphabet: c (dental click), q (palatal click), and x (lateral click) (see Appendix F). N|uu is a critically endangered "non-Bantu click language" in Southern Africa, belonging to the !UI language family of which only N|uu remains (Shah & Brenzinger, 2016:7-10). In addition to the 26 letters of the Roman alphabet, N|uu makes use of the following characters to represent its clicks: ʘ, |, ’, !, ǂ, and ǁ (Shah & Brenzinger, 2016:79-80).

In N|uu, the word **ǁx'â** is the translation equivalent of *belly* and *stomach* ("ǁx'â", 2016:80,115). The encoding of the non-alphabetic characters in the resource identifier (and thus the URI) has been identified as a potential challenge. Furthermore, the associated language code may also present a challenge should other dialects be encoded. Before the encoding of non-alphabetic characters is discussed, the language code and language tagging of N|uu is briefly addressed.

5.2.1 The language code

In this case study, a language code is used in several locations:

- in a lexical entry: for the language-tagging of literals;
- in a lexical concept: each definition and example in the N|uu language would be language-tagged;
- in the resource identifier, which forms part of the URI, for both a lexicon and a lexical entry.

N‖ng is the name of the dialect cluster of which N|uu is the Western variety, and ‖'Au, the Eastern variety (Güldemann, 2014:17; Van Der Merwe, 2015). The ISO 639-3 language code for N‖ng is "ngh", shared by the N|uu and ‖'Au dialects (SIL International [SIL], 2017a). However, according to the archival Khoisan "doculects" discussed by Güldermann (2014:16), the corpora from the Western variety of N‖ng are unknown or rejected by speakers of the Eastern variety of N‖ng. Should the now-extinct ‖'Au dialect be encoded using Ontolex-Lemon, it may be necessary to distinguish between the two dialects. In MultiTree, a library of language relationships hosted by The Linguist List, the codes for N|uu and ‖'Au are "ngh-nuu" and "ngh-aun" respectively ("N|u of N‖ng (ngh)", n.d.; "|'Auni of N‖ng (ngh)", n.d.). Both are documented for "Private Use", however their syntax does not meet the requirement defined by IETF's BCP 47, where the private use portion of the tag must be prepended with "x-" (W3C, 2014b; Phillips & Davis, 2009:3-4). For the latter portion of MultiTree's codes, namely "nuu" and "aun", both are also pre-existing language codes, the former for the language Ngbundu, and the latter for Molmo One ("Language-subtag-registry", 2018), and although point 2.2.7.5 of BCP 47 (Phillips & Davis, 2009:17) states that use of "private use subtags is by private agreement only", unless explicitly defined prior to use, the use of MultiTree's language tags may be inadvertently misinterpreted when encoded in a URI. For this reason, the use of Glottolog, a comprehensive catalogue of the world's lesser-known languages maintained by the Max Planck Institute for the Science of Human History, is suggested, as their catalogue "assigns a unique and stable identifier (the Glottocode) to (in principle) all languoids, i.e. all families, languages, and dialects" (Hammarström, Forkel & Haspelmath, 2018).

Including the Glottocode with the language code for N‖ng ("Language: N/u", n.d.), the language tags for N|uu and ‖'Au can be defined as:

- N|uu: `ngh-x-nuuu1242`
- ‖'Au: `ngh-x-auni1243`

5.2.2 Encoding of the resource identifier

As described in Section 3.2.4, the resource identifier begins with the language code, so continuing with the lexical entry ǁ*x'â*, the resource identifier is defined as:

- `ngh_x_nuuu1242-n-ǁx'â`

The URI for `ngh_x_nuuu1242-n-ǁx'â` can be written as follows:

U1: `https://londisizwe.org/entry/ngh_x_nuuu1242-n-ǁx'â`

However, the URI protocol only allows the US-ASCII character set, thus **U1** contains incompatible characters, namely "ǁ", "'" and "â" (Berners-Lee, Fielding & Masinter, 1998:7). The Internationalized Resource Identifier (IRI) protocol extends the character set of URIs, thus **U1** is a valid IRI, but in order for it to be a valid URI, it will need to be percent-encoded (Duerst & Suignard, 2005:2,12-13):

U1: `https://londisizwe.org/entry/ngh_x_nuuu1242-n-%C7%81x%E2%80%99%C3%A2`

Using content negotiation, **U1** will resolve to **U3**, which is a written representation of the resource identified in **U1**, either as an HTML file or a file containing an RDF serialisation, with each named according to the resource identifier. However, depending on the configuration of the server, both the unencoded and encoded form of the resource identifier contain characters which may not be supported in filenames. To resolve this, the server configuration file may need to include a mapping from **U3** to a file with an allowed filename, or an alternative version of the resource identifier could be considered. Continuing with the theme of descriptive URIs described in Section 3.2.4 for lexical entries, the following rules could be considered for the resource identifier:

- strip reserved and unsafe characters from the resource identifier, and
- remove diacritics from any characters in the resource identifier.

When the same rules are applied to the lexical entry |*x'a* (meaning "hand") ("|x'a", 2016:100), a URI collision will result as both resource identifiers would be identical (W3C, 2004a). Not only does this impact the URI, as the same URI cannot be used to

identify two different resources, but the benefit of using a descriptive URI for lexical entries is negated due to the possible incorrect assignment of meaning by the human user when identifying a lexical entry on the basis of the resource identifier. Thus, the suggestion is to include the English translation equivalent in the resource identifier as well:

- `ngh_x_nuuu1242-n-xa_belly`

The URI would then be written as follows:

U1: `https://londisizwe.org/entry/ngh_x_nuuu1242-n-xa_belly`

5.2.3 Describing the concept: `000000008`

Code Example 5-1: Describing the concept 000000008

Where:

- Line 6: shows the language tagging of N|uu
- Line 9: shows the resource identifier for ǁ*x'â*, the translation equivalent of *belly*.

5.2.4 Conclusion

The research question, Q4, was addressed in this section:

Q4: *"How does one construct the LLDF for lexical entries which may have letters not typically represented by the Roman alphabet?"*

A discussion of the encoding of non-alphabetic characters, primarily in the resource identifier, which forms part of the URI, addressed this question.

5.3 Construction of an LLDF – an overview

In this section, the primary research question is addressed, namely:

Q0: *"How does one construct a framework for bilingual lexicographic resources, applying linguistic linked data principles?"*

The methodological guidelines for publishing linked data in the available literature were presented in Section 3.1, with a comment by the researcher that the guidelines from Vila-Suero et al. (2014:103-15) and Gracia and Vila-Suero (2015), both specific to the publication of bilingual and multilingual resources, could be subject to further refinement.

The methodology for the construction of an LLDF for bilingual lexicographic resources, as identified in the preceding chapters, as well as the subsequent generation and publication of the linked data, is thus refined by the researcher as follows:

1. Identify the use cases.
2. Select the model with which to describe the linguistic data in RDF in a principled way.
3. Identify the external resources to link to, as well as any other ontologies and vocabularies to use.
4. Identify a strategy for versioning.
5. Identify the RDF formats required.
6. Design the URI patterns.
7. Consider the lemmatisation approach (if modelling agglutinating languages).

8. Model a lexical entry, a lexicon, and a lexical concept, using the resources identified in Step 3.
9. Generate the data as well as any associated metadata in the required RDF formats, as per the modelling identified in Steps 7 and 8, using the URI patterns from Step 6, and employing the versioning strategy from Step 4.
10. Publish the RDF data.
11. Return to Step 8 and repeat as required.

This methodology serves as a simplification of the process of the conversion of a bounded and static lexicographic resource to an unbounded and evolving framework, described using linguistic linked data principles. Although it must be noted that the methodology is not presented as 'best practice', the construction of the framework, as well as the generation and publication of the RDF data, is intended to be applied to other lexicographic resources as well, thereby contributing to the building of repeatable frameworks within the domain of LLD.

Each step in the methodology is expanded on, with pointers to the relevant sections in the preceding chapters:

Step 1: Identify the use cases

As described in Section 2.1.1, a single case design was adopted, using EXDN, a dictionary described in Section 1.2.1. The use case that results from this single case design is as follows:

- The representation of lexical entries in RDF:
  - where English is the source language and isiXhosa is the target language,
  - representation is unidirectional only,
  - representation is diachronic (as opposed to EXDN, which is synchronic).

However, as outlined in Section 2.1.1, because the context of the original lexical entries in EXDN was expanded to include other languages in South Africa, the scope consequently changed:

- The representation of lexical entries in RDF:

- where English and isiXhosa are two of the languages,
- representation should be bidirectional,
- the representation of languages with click consonants which do not use the Roman alphabet is considered,
- representation is diachronic.

Step 2: Select the model

The selection of an appropriate model for representing LLD was detailed in Section 2.2, with the modelling requirements identified in Section 2.3, and the selected model, Ontolex-Lemon, detailed in Section 2.4. The use cases identified in Step 1 would have to be supported by the model.

Step 3: Identify the external resources to link to

The external resources to link to (a Linked Data principle by Berners-Lee, discussed in Section 3.1) were listed in Section 2.6.4, as well as other ontologies and vocabularies used to describe the data.

Step 4: Identify the versioning strategy

The versioning strategy for lexical entries in a RDF repository was discussed in Section 3.8.1.

Step 5: Identify the RDF formats

The RDF formats were identified in Section 1.6, with a detailed description of the formats provided in Appendix B.

Step 6: Design the URI patterns

The URI patterns for six URI use cases were designed in Section 3.2.3.

Step 7: Consider the lemmatisation approach

The lemmatisation approach for agglutinating languages was discussed in Section 4.3.

Step 8: Model a lexical entry, a lexicon, and a lexical concept

The namespaces for the ontologies and vocabularies identified in Section 2.6.4, as used in the representation of RDF data by this study, were listed in Section 2.6.6. Modelling of a lexical entry, the lexicon, translation equivalence, and provenance were described in Sections 3.4 to 3.8. The identification of the use cases for modelling key characteristics of the lexical entries was described in Section 4.2 to 4.4. Modelling of a lexical concept was described in Section 4.4.5.

Step 9: Generate the RDF data

The generation of RDF data was described in Section 3.9.

Step 10: Publish the RDF data

The publication of RDF data was described in Section 3.9.

In order to account for change, the methodology is intended to be iterative, with Steps 8-10 accounting for the greatest change.

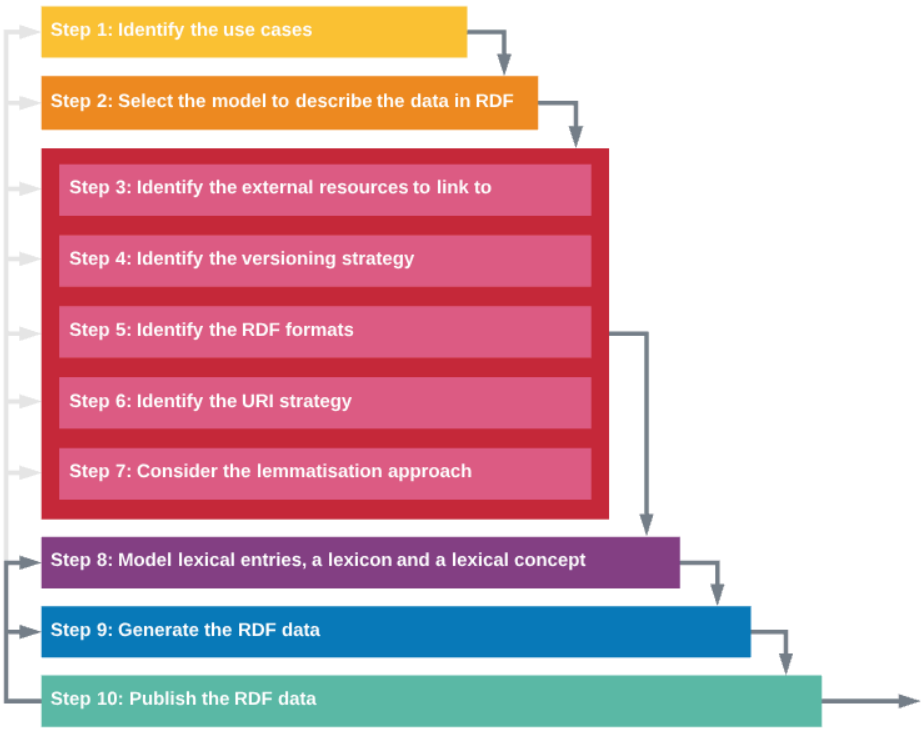


Figure 5-1: Methodological guidelines for an LLDF

However, as shown in Figure 5-1, strict sequentiality is not required: Steps 3-7 can be reordered as necessary.

To conclude the discussion of the question Q0, the identified sub-objectives and research questions, Q1 – Q4, and findings thereof, are reviewed in accordance with the model, Ontolex-Lemon, and the literature available for other lexicographic resources which made use of the same model (or its predecessor, *lemon*). Similar studies include ABD (Gracia et al., 2018; Bosque-Gil, Gracia & Montiel-Ponsoda, 2017), AQAM (Khalfi, Nahli & Zarghili, 2016), DEAF (Tittel & Chiarcos, 2018) and Zhishi.me (Fang et al., 2016), with ABD, AQAM and DEAF discussed in more detail in Section 2.5. Zhishi.me is a Chinese dataset in the LLD cloud, and the Zhishi.lemon dataset was built which "constitutes the lexical realisation of Zhishi.me", using *lemon* as the model (Fang et al., 2016:48).

5.3.1 (Q1) How does one construct the LLDF so as to allow for extensibility from a bilingual to a multilingual resource?

If only the single case design was adopted, then the model, using its *vartrans* module, would have been sufficient to describe the unidirectional English→isiXhosa language pair, as was done for each of the Apertium dictionaries, of which twenty-two were transformed to RDF (Gracia et al., 2018:6). However, by expanding the scope to describe a bidirectional English⇔isiXhosa language pair, the model and its *vartrans* module could still be used, although redundancy is introduced as the triple for each unidirectional translation equivalence relationship would have to be declared (effectively reversing the bilingual entry); the researcher considers it insufficient to declare a single translation equivalence relationship (see Use Case 2, discussed in Section 3.7) for a source and target language of a language pair and then assume the inverse of that relation. By introducing a third language to the resource, for example by digitising a monolingual dictionary from the same domain, making the combined resource multilingual, even more redundancy would result, as each translation equivalence relationship would have to be declared between each language pair.

Use Case 1 of the *vartrans* module (also discussed in Section 3.7) recommended the use of an ontological entity as a shared reference to indicate translation equivalence. However, context and cotext are important; thus, while two lexical entries may be ontologically equivalent, they share only partial equivalence. Using an ontological entity to define a lexical entry (the *semantics by reference* principle employed by the model), or to show equivalence between lexical entries from different lexicons (Use Case 1 of the *vartrans* module of the model) came in for criticism by Hirst (2014:5):

> … the utility of ontologies for interpreting linguistic information is thereby limited, and so, conversely, is the ability of lexicons to express ontological concepts. This leads to practical limitations on models of lexicons for ontologies, such as McCrae et al.'s (2012) *lemon* model, that put an emphasis on *interchangeability* – the idea that one ontology can have many different lexicons, for example, for different languages or dialects. This wrongly assumes that *translation-equivalent words* have identical meanings.

The solution was to construct a lexical concept as a shared conceptualisation, in which both the context and cotexts could be defined, and then to indicate the senses which can be lexicalised by this lexical concept, using `ontolex:lexicalizedSense`. However, this still proved insufficient as there are lexical gaps between languages, and very rarely is there full equivalence, with Hirst defining translation-equivalent words as "cross-lingual near-synonyms" (2014:5). This was demonstrated by modelling the English terms *abdomen*, *stomach*, and *belly*, all of which are equivalent to the isiXhosa term *isisu*. To model the English terms, three lexical concepts need to be defined (Hirst, 2014:6), each with the context and cotext of use indicated within the lexical concept. In the examples which described lexical concepts in the previous chapters, a lexical concept was declared to be `owl:sameAs` a WordNet synset available in RDF. When one considers the WordNet synset: {*abdomen*, *stomach*, *belly*, *venter*}, it is not entirely accurate to describe the lexical concept as `owl:sameAs` to the synset. The lexical concept can only be considered the same if context and cotext are excluded from the lexical concept, and even then, it is still not *the same*; rather, the lexical concept *is expressed by* a member of the synset (Fellbaum, 2006:665). Although the lexical concepts will become quite fine-grained, particularly as additional languages are added to the resource, including a reference to the member of the synset which expresses the lexical concept allows more coarsely-grained non-hierarchical near-synonyms to be associated with the lexical

concept (Hirst, 2014:6), thus serving as a way to categorise (or ground) the lexical concept.

To model this expression, for each `ontolex:lexicalizedSense` of a language in the lexical concept, a reference to the member of the synset of that language needs to be included. If a synset is not available for one or more of the languages in the lexical concept, then the lemma of the lexical entry which has a lexicalised sense in the lexical concept serves as a singular member of a newly-created synset for that language.

The revised lexical concept is demonstrated in Code Example 5-2.

Code Example 5-2: The revised lexical concept

Where:

- Line 9: indicates the reference to the respective member in the synset for each lexicalised sense. For the second synset member `<https://wn.londisizwe.org/xh/000000001-n#isisu>`, as the synset has not been created yet, this URI (compiled using the pattern of PWN's URIs as guidance) serves to identify only and is not dereferenceable.

The outcome of this is that Londisizwe Concepts for Senses, described in Section 2.6.5 and intended as a standalone inventory of lexical concepts, was erroneous. Due to the lexicalisation of senses within a lexical concept, lexical concepts could never truly be

separated from the primary data layer within the resource and therefore no distinction needs to be made in this regard. However, the WordNet for isiXhosa synsets is distinct from the resource. Therefore it should be a standalone resource, hence the differing format in the `{domain}` portion of the URI: `wn.londisizwe.org`, to that of the URI for a lexical entry, where `{domain}` is of the form `londisizwe.org` only. It must be noted that the WordNet for isiXhosa synsets is not just limited to isiXhosa; any lemma of a lexical entry of other languages for which a sense is lexicalised in a lexical concept could be represented in the WordNet, should there not be an appropriate synset in PWN.

5.3.2 (Q2) How does one construct the LLDF to allow for change, tracking provenance of each change?

As mentioned in Section 3.8.1, there is little information in the literature related to the management of provenance and versioning for language resources in the domain of LLD. In Ontolex-Lemon, the Lime module provides some metadata, and the metadata relevant to this study was described in Section 3.5.1. Due to this gap in the literature, the researcher presented a paper based on this study, entitled "Managing Provenance and Versioning for an (Evolving) Dictionary in Linked Data Format", at the 6th Workshop on Linked Data in Linguistics, co-located at the Language Resources and Evaluation Conference (LREC 2018) (Gillis-Webber, 2018b).

5.3.3 (Q3) How does one construct the LLDF for translation equivalents, which may have a differing lemmatisation approach for nouns and verbs?

Arabic words are derived from the root, and in AQAM, although lexical entries are ordered by the root, the lemmatisation approach was not discussed (Khalfi, Nahli & Zarghili, 2016:325). The root was included in code examples; however, these code examples appear to be XML prior to their conversion to *lemon*. In the appendix, two example lexical entries were encoded using *lemon*, but both excluded the representation of the root (Khalfi, Nahli & Zarghili, 2016:330).

The lemmatisation approach was not discussed in any of the other similar studies either; however, when considering translation in both ABD and Zhishi.me, translation relations rely on language pairs (Gracia et al., 2018:2; Fang et al., 2016:51). Although not

demonstrated in the examples of ABD and Zhishi.me, a language pair could have a differing lemmatisation approach, with the source language, for example *English*, being a standalone word and the target language, for example *isiXhosa*, being a stem. Therefore the language pair may be semantically equivalent but not lexically equivalent.

Continuing with the discussion of language pairs, as mentioned in Section 3.7.3, Gracia et al. (2018:7-8) talk of a pivot language to derive indirect translations. However, this would only be possible if there is full equivalence within a language pair. Where there is no full equivalence, using a pivot language may result in inaccuracies. Gleason has given an example in which the term *chichena* in Shona refers to the spectrum of colours identified in English as both *green* and *yellow* (1961:4).

Fang et al. (2016:51) went on to say that

> translation relations can be inferred between terms in different languages when they refer to the same ontology entity. These lexical senses with an equivalent ontology reference have been regarded as a translation pair to be modelled.

However, this was shown to be incorrect in Section 3.7.3: senses in a language pair may be ontologically equivalent but share only partial equivalence. If "ontology entity" had to be replaced with "lexical concept", then Fang et al.'s statement would be more accurate.

5.3.4 (Q4) How does one construct the LLDF for lexical entries which may have letters not typically represented by the Roman alphabet?

Working with the open-source dictionary of N|uu resulted in the identification of two issues that would not have been considered if only working with the EXDN dataset, namely the language codes for dialects, and characters with diacritics. No language codes for N|uu and ǁ'Au are available in ISO 639, with the result that the researcher created a 'compiled' language tag for each, comprising (a) the ISO 639-3 code for Nǁng, the dialect cluster of N|uu and ǁ'Au, (b) followed by "x-", which indicates private usage, and (c) followed by the use of Glottolog for the two dialects. In the similar study of DEAF, language codes in ISO 639 was identified as an issue for Old French dialects, with Glottolog as an alternative coming in for criticism by Tittel and Chiarcos as "not

appropriate for the needs of philologists; as an example, it conflates diachronic and dialectal criteria within a single hierarchy" (2018:7). The solution suggested by Tittel and Chiarcos is to define the language code 'fro' "as a macrolanguage and to register the Old French dialects as varieties associated to 'fro'" (2018:8).

Although the language identification requirement is not a fault of Ontolex-Lemon, when modelling data from under-resourced or little-known languages, dialects, or the diachronicity of a language, if the language has not been identified in ISO 639, there is going to have to be shared agreement of the 'compiled' language tags that result.

Considering another language, SASL (although not accorded official language status, it falls under the mandate of PanSALB with the Khoisan languages) has considerable variation in the lexicon, and according to Van Niekerk, as a result of South Africa's history of segregation, "the result was not only great variety in the SASL lexicon, but reduced contact between these schools which caused the language variety to become entrenched within the Deaf community" (National Institute for the Deaf, 2016), with a variation confined "to the community around each school" (Van Niekerk, personal communication 2018, July 10).

The language code for SASL, described in ISO 639-3, is "sfs" (SIL, 2017b). Glottolog has also assigned an identifier to SASL, but unlike that for N‖ng, its varieties have not been identified ("Language: South African …", n.d.). Ethnologue has identified twenty-nine schools for the Deaf, and a standardised variety is also promoted by DeafSA, the Deaf Federation of South Africa (SIL, 2018).

Although a solution was devised using a *privateuse* sub-tag, and Tittel and Chiarcos suggested defining the language code as a macrolanguage, with the dialects registered as varieties with ISO 639, the challenge with both solutions in the instance of SASL is that there is not necessarily shared agreement regarding the varieties of SASL. With the current language tagging solution offered by W3C, due to the inability to use a URI to identify the tag, shared agreement amongst resources in the LL(O)D cloud is not possible.

Language tags take the form: *language-extlang-script-region-variant-extension-privateuse*, where *language* is a language code from ISO 639 and the remaining sections are sub-tags, each presented in a different format so as to identify them; for example, *extlang* is capitalised and *script* is written in sentence case (W3C, 2014b). Although the language tags can be encoded with valuable information, when querying the dataset using SPARQL, knowledge of the language tags used in the dataset will be needed. It may therefore be better to model any dialects, scripts, variation and other language identification information in RDF as well, so that it can be queried in SPARQL. The Ontolex-Lemon specification does not provide for this information, which would probably need to be represented in both the lexical entry and the lexical concept. Describing the data in a principled way will lead to increased agreement as external resources will be able to refer to these same triples which serve as a language identifier, using URIs.

To conclude the discussion of Q0 and its sub-objectives Q1 – Q4, in the similar studies of DEAF and ABD, additional issues were identified that cannot (currently) be resolved in Ontolex-Lemon:

- the ordering of senses (Bosque-Gil, Gracia & Montiel-Ponsoda, 2017:9), and the ordering of sub-senses in relation to main senses (Tittel & Chiarcos, 2018:8);
- the modelling of headwords that can take a different POS (Bosque-Gil, Gracia & Montiel-Ponsoda, 2017:9).

5.4 Future work

The following topics have been identified for further research:

- Create the synsets for the URIs in the lexical concepts which are not currently dereferenceable and, where possible, indicate the semantic relations between the synsets. This WordNet should be linked data-native.
- Consider an alternative solution to language tags (although not intended to replace language tagging), so as to render language identification more accurate in the

LL(O)D cloud; this in turn can lead to shared agreement between lexical resources, important in a MSW.

- Ontolex-Lemon states when referring to its purpose (Cimiano, McCrae & Buitelaar, 2016):

> It is not a formal model of semantics but a model of lexicography. The model is not supposed to be used to define an ontology and instead assumes that there is a given ontology in some ontology language that is to be linked to a lexicon that expresses how the classes, properties and individuals defined in the ontology are lexicalized.

For this study, a formal ontology was not included in the framework; however, Khalfi, Nahli and Zarghili talk of using SUMO "where concepts are defined with machine-interpretable semantics in first order logic" (2016:328). In order to contribute to the vision of the MSW (discussed in Section 1.5) beyond the framework described in this study, formal ontology will have to play a bigger role. As a starting point, the linking to a higher-level ontology, such as Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) or Basic Formal Ontology (BFO), in the lexical concept should be considered.

5.5 Conclusion

The output of the construction of an LLDF using Ontolex-Lemon to describe a digitised bilingual lexicographic resource is a multilingual, machine-interoperable lexical resource in Linked Data format. To conclude this study, this lexical resource is considered through the lens of 'context' as defined by León-Araúz and Faber (2014:31) as "the parts of a written or spoken statement that precede or follow a specific word or phrase and which can influence its meaning or effect. It is also the situation, events or information that are related to something and which help a person to understand it."

Ontolex-Lemon's purpose is "to support linguistic grounding of a given ontology", where "the semantics of a lexical entry is expressed by reference to an individual, class or property defined in the ontology (Cimiano, McCrae & Buitelaar, 2016). However, as demonstrated with the lexical entries *breath* and *breathing*, there is not always an ontological entity to support the semantics of a lexical entry. According to Hirst, the

semantics by reference principle "leads to an inflexible and limiting view of word senses" (2014:5). Hirst goes on to say that the "problems are compounded when we add multilinguality as an element", where the assumption that translation-equivalent words have identical meaning and the notion of an interchangeable lexicon are simply incorrect (2014:5). In the field of lexicography, Zgusta talks of "equivalent lexical units with identical multiple meaning in both languages, and with precisely the same lexical meaning" as a "real rarissimum" (1971:296).

Staying within lexicography, Gouws and Prinsloo (2005:162) talk of communicative equivalence, saying:

> [it] can only achieved if the treatment is not restricted to a listing of equivalents but if these equivalents are complemented by context and cotext entries that can help the user to choose the correct equivalent for a given occurrence of the source language item and to use this equivalent in a proper way.

In a separate paper, Gouws describes bilingual dictionaries as "aids in interlingual translations" and "their main function is not a transfer of meaning", but rather "to endeavour to reach communicative equivalence" (1996:16).

However, as pointed out by Hirst, "a dictionary is intended for use by humans, and its style and format are unsuitable for computational use in a text or natural language processing system". He goes on to say that definitions are written in natural language, but "computational applications that use word meanings usually require a more-formal representation of knowledge" (2009:2). Zgusta, describing equivalence between entries in a Chinese-German dictionary, talks of indicating the restrictions in a particular context, and this restriction might need to only be indicated in one direction (1971:316-317).

When one considers the lexical concepts of this study, where context is defined using natural language within the lexical concept and where it is expected that there will rarely be more than one lexicalised sense within a concept, this means there will only be one synset, which in turn means the discovery of synsets of other languages which share near-synonyms will not be possible. To address this, the WordNet discussed as future work in Section 5.4 can be encoded in a way that allows it to contribute to the Collaborative Interlingual Index (CILI), an index of WordNets, with each WordNet

sharing a common format in three forms: XML, JSON and RDF, where "linking is made between the synsets" (McCrae et al., 2017:591). Each synset is assigned an ILI identifier, with each identifier associated with its own URL (McCrae et al., 2017:593). By indicating the ILI identifier in the lexical concept, it is possible for interlingual synsets to be discovered.

Ontolex-Lemon provides the means to model restrictions in a lexical entry or a sense therein, such as identifying the diaphrasic variant that the term is typically used in, for example H_2O is used within a scientific context. However, modelling restrictions on context and domain for translation equivalence is not so easily resolved. Depending on the use case, one could argue for the restriction to be modelled within the ontology, the lexical concept, the synset, or even a combination thereof. The end-result is a messy solution, one that is not neatly defined nor easily contained. Harking back to the comment by Hirst in Section 3.6.5, "there is, in practice, no clean separation between the conceptual and the linguistic" (2014:7).

To conclude this study, as mentioned in Section 2.1.1, the aims of the case study approach were (1) to generate theory, (2) to test theory, and (3) to provide description. For the purpose of this study, the three aims of the case study approach were achieved (Eisenhardt, 1989:535; Nazari, 2010:180):

- the phenomenon and its context were explored, with description provided:
  the dictionary, its digitisation, and its retrodigitisation to a complex digital object was described, and the encoding of the complex digital object was detailed, using RDF, adhering to Linked Data principles, and describing the data in a principled way using a model, Ontolex-Lemon;
- the theoretical perspective was tested within the specific context:
  abstractions of the key characteristics of lexical entries were modelled using Ontolex-Lemon;
- and a model was generated, by building on the existing theory:
  the methodological guidelines for publishing Linked Data of language resources was evaluated and refined, a research gap on provenance and versioning for language resources was filled by generating a new model, and the model on lexical concepts

was expanded upon, with the inclusion of synsets, where a member of a synset expresses a lexical concept within a multilingual context. From this, an LLDF has been (re-)defined (Gillis-Webber, 2018a:4):

> as a framework that:
>
> 1. describes data in RDF,
> 2. uses a model designed for the representation of linguistic information,
> 3. adheres to Linked Data principles, and
> 4. supports versioning.

Ontolex-Lemon is intended for an ontology where the lexicon can be changed (McCrae et al., 2017:1), effectively giving the ontology multilingual `rdfs:label`s. However achieving communicative equivalence within an ontology is not possible as it requires action by a human to make a selection from context and cotexts, and to interpret any restrictions on a context, so one should focus less on equivalence and focus more on *intelligibility* by way of shared conceptualisation.

In the words of McArthur (1986:11), "a printed and bound dictionary, for example, is like a fossil; the moment it is complete and published, it is dated and rendered imperfect by the continuing flow of the language beyond what it has described." To take a lexicographic resource from a state that is bounded, static and reliant on humans to infer meaning, to an LLDF that is unbounded, evolving, and reliant on a machine to infer meaning, is not without difficulty. Hirst (2014:12) maintains that "the future of semantic representations for the Multilingual Semantic Web is likely to lie in imperfect nonsymbolic methods that work well enough in practice for most situations." McCrae et al. (2017:1) talk of switching the lexicon, saying "we can easily switch an ontology from one language to another by changing its lexicon", but if by doing so, and if as a result only achieving *intelligibility* is insufficient, maybe the solution is to rather switch the ontology?

References

2nd summer datathon on linguistic linked open data (SD-LLOD-17). Cercedilla, Spain, 26-30 June 2017. Available: http://datathon2017.retele.linkeddata.es/ [2017, December 18].

"Abdomen". *English-Xhosa dictionary for nurses.* 1935. 2nd ed. Lovedale, South Africa: Lovedale Press.

"abdomen". *Oxford English Xhosa dictionary.* 2013. 25th ed. Cape Town: Oxford University Press Southern Africa (Pty) Ltd.

"abdomen • stomach • belly • venter". n.d. *BabelNet.* Available: http://babelnet.org/synset?word=bn:00000249n&details=1&lang=EN&orig=abdomen [2017, December 30].

Abele, A., McCrae, J.P., Buitelaar, P., Jentzsch, A. & Cyganiak, R. 2017. *Linking open data cloud diagram*. Available: http://lod-cloud.net/ [2017, November 4].

Alexander, K., Cyganiak, R., Hausenblas, M. & Zhao, J. 2011. *Describing linked datasets with the VoID vocabulary: W3C Interest Group Note 03 March 2011*. Available: https://www.w3.org/TR/void/ [2018, January 10].

Anfara, V.A. & Mertz, N.T. 2006. Conclusion: coming full circle. In *Theoretical frameworks in qualitative research*, 189-196. V.A. Anfara, Jr. & N.T. Mertz, Eds. California: Sage Publications, Inc. Pages.

Anfara, V.A. & Mertz, N.T. 2006. Introduction. In *Theoretical frameworks in qualitative research*, xiii-xxxii. V.A. Anfara, Jr. & N.T. Mertz, Eds. California: Sage Publications, Inc. Pages.

application/rdf+xml. 2004. Available: https://www.iana.org/assignments/media-types/application/rdf+xml [2017, December 5].

"Aqua or Aq". *English-Xhosa dictionary for nurses.* 1935. 2nd ed. Lovedale, South Africa: Lovedale Press.

Archer, P., Goedertier, S. & Loutas, N. 2012. *D7.1.3 – Study on persistent URIs, with identification of best practices and recommendations on the topic for the MSs and the EC.* Available: https://joinup.ec.europa.eu/sites/default/files/document/2013-02/D7.1.3%20-%20Study%20on%20persistent%20URIs.pdf [2017, December 26].

Aronin, L. & Singleton, D. 2012. *Multilingualism*. Amsterdam: John Benjamins Publishing Company.

Arp, R., Smith, B. & Spear, A.D. 2015. *Building ontologies with Basic Formal Ontology.* Cambridge, Massachusetts: The MIT Press.

Atkins, B.T.S. & Rundell, M. 2008. *The Oxford guide to practical lexicography*. Oxford: Oxford University Press.

BabelNet. n.d. *About*. Available: http://babelnet.org/about [2017, December 30].

"belly". *Oxford English Xhosa dictionary.* 2013. 25th ed. Cape Town: Oxford University Press Southern Africa (Pty) Ltd.

Benefits. 2011. Available: https://www.w3.org/2005/Incubator/lld/wiki/Benefits [2017, November 4].

Berners-Lee, T. 2006. *Linked data*. Available: https://www.w3.org/DesignIssues/LinkedData.html [2017, April 15].

Berners-Lee, T. 2009. *Tim Berners-Lee: the next web* [Video file]. Available: http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html [2017, April 15].

Berners-Lee, T., Fielding, R., Masinter, L. 1998. *Uniform resource identifiers (URI): generic syntax*. Available: https://www.ietf.org/rfc/rfc2396.txt [2018, June 25].

Berners-Lee, T., Hendler, J. & Lassila, O. 2001. The semantic web. *Scientific American.* 284(5):34-43. Available: http://www.jstor.org/stable/26059207 [2017, November 4].

Best practices for multilingual linked open data community group: W3C community and business groups. 2017. Available: https://www.w3.org/community/bpmlod/ [2017, December 25].

Bizer, C., Heath, T. & Berners-Lee, T. n.d. *Linked data – the story so far.* Available: http://tomheath.com/papers/__bizer-heath-berners-lee-ijswis-linked-data.pdf [2017, December 18].

Bosque-Gil, J. 2017. Linked data and dictionaries [Seminar]. 2nd Summer Datathon on Linguistic Linked Open Data. 27 June 2017.

Bosque-Gil, J., Gracia, J., Aguado-de-Cea, G. & Montiel-Ponsoda, E. 2015. Applying the OntoLex model to a multilingual terminological resource. In *The semantic web: ESWC 2015 satellite events.* F. Gandon, C. Guéret, S. Villata, J. Breslin, C. Faron-Zucker & A. Zimmerman, Eds. 283-294.

Bosque-Gil, J., Gracia, J. & Gómez-Pérez, A. 2016. Linked data in lexicography. *Kernerman Dictionary News*. 24:19-24.

Bosque-Gil, J., Gracia, J. & Montiel-Ponsoda, E. 2017. Towards a module for lexicography in OntoLex. *Proceedings of the LDK 2017 Workshops: 1st Workshop on the OntoLex Model (OntoLex-2017), Shared Task on Translation Inference Across Dictionaries & Challenges for Wordnets*. Galway, Ireland, 18 June 2017. Available: http://ceur-ws.org/Vol-1899/OntoLex_2017_paper_5.pdf [2018, January 16].

Bouda, P. & Cysouw, M. 2012. Treating dictionaries as a linked-data corpus. In *Linked Data in Linguistics*. C. Chiarcos, S. Nordhoff & S. Hellman, Eds. Heidelberg: Springer. 15-24.

"Breath". *English-Xhosa dictionary for nurses.* 1935. 2nd ed. Lovedale, South Africa: Lovedale Press.

Brickley, D. & Miller, L. 2014. *FOAF vocabulary specification 0.99: namespace document 14 January 2014 - Paddington edition.* Available: http://xmlns.com/foaf/spec/ [2018, January 10].

Buitelaar, P., Choi, K., Cimiano, P. & Hovy, E.H. Eds. 2012. The multilingual semantic web (Dagstuhl Seminar 12362). *Dagstuhl Reports*. 2(9):15-94.

Bussmann, H. 1996. *Routledge dictionary of language and linguistics.* Translated by G. Trauth & K. Kazzazi. London: Routledge.

"Change of life". *English-Xhosa dictionary for nurses.* 1935. 2nd ed. Lovedale, South Africa: Lovedale Press.

Chavula, C. & Keet, C.M. 2014. Is lemon sufficient for building multilingual ontologies for bantu languages? *Proceedings of the 11th OWL: Experiences and Directions Workshop (OWLED'14)*. Riva del Garda, Italy, 17–18 October, 2014. Available: http://ceur-ws.org/Vol-1265/owled2014_submission_10.pdf [2018, January 5].

Chowdhury, G. 2015. Management of cultural heritage information: policies and practices. In *Cultural heritage information: access and management.* I. Ruthven & G.G. Chowdhury, Eds. London: Facet Publishing. 37-62.

Cimiano, P., Buitelaar, P., McCrae, J. & Sintek, M. 2010. LexInfo: a declarative model for the lexicon-ontology interface. *Web Semantics: Science, Services and Agents on the World Wide Web.* 9:29-51.

Cimiano, P., McCrae, J.P. & Buitelaar, P. Eds. 2016. *Lexicon model for ontologies: community report, 10 May 2016*. Ontology-Lexicon Community Group under the W3C Community Final Specification Agreement (FSA). Available: https://www.w3.org/2016/05/ontolex/ [2017, October 27].

Cimiano, P., McCrae, J., Buitelaar, P. & Montiel-Ponsoda, E. 2012. *On the role of senses in the ontology-lexicon.* Available: https://lists.w3.org/Archives/Public/public-ontolex/2012Jun/att-0009/senses.pdf [2018, January 28].

Converting BabelNet as linguistic linked data. 2014. Available: https://www.w3.org/community/bpmlod/wiki/Converting_BabelNet_as_Linguistic_Linked_Data [2017, December 5].

Copyright Act, No. 98 of 1978, as amended. 2017. *Government gazette.* 5 July 2016. Government notice no. 40121. Cape Town: Government Printer. Available: https://www.gov.za/sites/default/files/B13-2017_Copyright_170516.pdf [2017, November 5].

Coyle, K. 2012. Chapter 2: Semantic web and linked data. *ALA Tech Source*. 4:10-14. DOI:10.5860/ltr.48n4.

Crystal, D. 1997. *A dictionary of linguistics & phonetics.* 4th Ed. Oxford: Blackwell Publishing.

Crystal, D. 2003. *A dictionary of linguistics & phonetics.* 5th Ed. Cambridge, Massachusetts: Blackwell Publishing.

Crystal, D. 2010. *The Cambridge encyclopedia of language.* Cambridge: Cambridge University Press.

DBpedia. 2018. *Learn about DBpedia: about*. Available: https://wiki.dbpedia.org/about [2018, January 1].

De Melo, G. 2015. Lexvo.org: language related information for the linguistic linked data cloud. *Semantic Web*. 6(4):393-400. Available: http://www.semantic-web-journal.net/system/files/swj521.pdf [2018, February 1].

De Rooij, S., Beek, W., Bloem, P., van Harmelen, F. & Schlobach, S. 2016. Are names meaningful? Quantifying social meaning on the semantic web. In *The Semantic Web: ISWC 2016.* P. Groth, E. Simperl, A. Gray, M. Sabou, M. Krötzsch, F. Lecue, F. Flöck & Y. Gil, Eds. 184-199.

De Schryver, G-M. 2010. Revolutionizing Bantu lexicography – a Zulu case study. *Lexikos.* 20:161-201.

The Department of Justice and Constitutional Development. 2017. *Founding provisions: chapter 1, section 1-6*. Available: http://www.justice.gov.za/legislation/constitution/chp01.html [2018, July 24].

Di Maio, P. 2015. Linked data beyond libraries. In *Linked data and user interaction.* H.F. Cervone, L.G. Svensson, Eds. Berlin: Walter de Gruyter GmbH.

Diekema, A.R. & Eccles, E. 2012. Multilinguality in the digital library: a review. *The Electronic Library*. 30(2):165-181. DOI: 10.1108/02640471211221313.

Doke, C.M. 1954. *The Southern Bantu languages.* London: International African Institute.

Dublin Core Metadata Initiative. 2012. *DCMI metadata terms*. Available: http://dublincore.org/documents/dcmi-terms/ [2018, January 10].

Duerst, M. & Suignard, M. 2005. *Internationalized resource identifiers (IRIs).* Available: https://tools.ietf.org/html/rfc3987 [2018, June 25].

Dydra. n.d. Import your data [Password protected]. Available: https://dydra.com/frangipaniza/londisizwe-org/@import [2018, January 3].

Dydra. 2011. Available: https://www.w3.org/2001/sw/wiki/Dydra [2018, January 3].

Eckart, K., Riester, A. & Schweitzer, K. 2012. A discourse information radio news database for linguistic analysis. In *Linked Data in Linguistics*. C. Chiarcos, S. Nordhoff & S. Hellman, Eds. Heidelberg: Springer. 65-76.

Ehrmann, M., Cecconi, F., Vannella, D., McCrae, J., Cimiano, P. & Navigli, R. 2014. Representing multilingual data as linked data: the case of BabelNet 2.0. *Proceedings of LREC 2014, Ninth International Conference on Language Resources and Evaluation*. 26-31 May 2014. Available: http://www.lrec-conf.org/proceedings/lrec2014/pdf/810_Paper.pdf [2017, December 5].

Eisenhardt, K.M. 1989. Building theories from case study research. *Academy of Management Review.* 14(4):532-550. Available: https://www.jstor.org/stable/258557 [2017, December 17].

Eisenhardt, K.M. & Graebner, M.E. 2007. Theory building from cases: opportunities and challenges. *Academy of Management Journal.* 50(1):25-32.

ELRA. 2015. *What is a language resource?* Available: http://www.elra.info/en/about/what-language-resource/ [2017, November 1].

English-Xhosa / Xhosa-English Dictionary. 2014. Cape Town: Pharos Dictionaries.

Espinoza, M., Gómez-Pérez, A. & Montiel-Ponsoda, E. 2009. Multilingual and localization support for ontologies. In *The Semantic Web: Research and Applications: Proceedings of the 6th European Semantic Web Conference, ESWC 2009.* 31 May-4 June 2009. L. Aroyo, P. Traverso, F. Ciravegna, P. Cimiano, T. Heath, E. Hyvönen, R. Mizoguchi, E. Oren, M. Sabou, & E. Simperl, Eds. 821-825.

European Commission. 2011. *Digital Agenda: recommendation on the digitisation of cultural material and its preservation on line – frequently asked questions*. Available: http://europa.eu/rapid/press-release_MEMO-11-745_en.htm [2017, March 29].

Faab, G., Bosch, S.E. & Gouws, R.H. 2014. A general lexicographic model for a typological variety of dictionaries in African languages. *Lexikos*. 24:94-115.

Fang, Z., Wang, H., Gracia, J., Bosque-Gil, J. & Ruan, T. 2016. Zhishi.lemon: on publishing Zhishi.me as linguistic linked open data. In *The Semantic Web: ISWC 2016.* P. Groth, E. Simperl, A. Gray, M. Sabou, M. Krötzsch, F. Lecue, F. Flöck & Y. Gil, Eds. 47-55.

Faniel, I.M. & Yakel, E. 2017. Practices do not make perfect: Disciplinary data sharing and reuse practices and their implications for repository data curation. In *Curating research data: practical strategies for your digital repository*. L.R. Johnston, Ed. Chicago: Association of College and Research Libraries. 103-126.

Fellbaum, C. 2006. WordNet(s). In *Encyclopedia of language & linguistics, second edition.* K. Brown, Ed. 13:665-670. Available: http://www.iaoa.org/isc2012/docs/encycloped.article.pdf [2018, August 5].

Flati, T., Moro, A., Matteis, L., Navigli, R. & Velardi, P. 2015. *Guidelines for linguistic linked data generation: multilingual dictionaries (Babelnet).* Available:

https://www.w3.org/2015/09/bpmlod-reports/multilingual-dictionaries/ [2017, December 27].

Francopoulo, G. & George, M. 2013. Model description. In *LMF – Lexical markup framework.* G. Francopoulo, Ed. London: ISTE Ltd.

"Fumigation". *English-Xhosa dictionary for nurses.* 1935. 2nd ed. Lovedale, South Africa: Lovedale Press.

Gillis-Webber, F. 2018. Conversion of the English-Xhosa Dictionary for Nurses to a linguistic linked data framework. *Information*. J.P. McCrae & J. Gracia, Eds. 9(11). Available: https://www.mdpi.com/2078-2489/9/11/274 [2018, November 15].

Gillis-Webber, F. 2018. Managing provenance and versioning for an (evolving) dictionary in linked data format. *Proceedings of the 6th Workshop on Linked Data in Linguistics: Towards Linguistic Data Science, Co-Located with LREC2018, Miyazaki, Japan.* 12 May 2018. Available: http://lrec-conf.org/workshops/lrec2018/W23/pdf/2_W23.pdf [2018, November 1].

Gleason, H.A. 1961. *An introduction to descriptive linguistics*. Revised ed. London: Holt, Rinehart and Winston.

Gomm, R., Hammersley, M. & Foster, P. 2000. Case study and generalization. In *Case study method.* R. Gomm, M. Hammersley & P. Foster, Eds. London: Sage Publications Ltd.

Google Cloud. 2018. *Translating text.* Available: https://cloud.google.com/translate/docs/translating-text [2018, November 24].

Gouws, R.H. 1996. Bilingual dictionaries and communicative equivalence for a multilingual society. *Lexikos.* 6:14-31.

Gouws, R.H. 2018. A dynamic lexicographic practice for diverse users and changing technologies. In *The dynamics of language: plenary and focus lectures from the 20th International Congress of Linguists*. Cape Town: UCT Press. 231-244.

Gouws, R.H. & Prinsloo, D.J. 2005. *Principles and practice of South African lexicography*. Stellenbosch: SUN MeDIA.

Gracia, J. 2017. Introduction to linked data for language resources [Practical session]. 2nd Summer Datathon on Linguistic Linked Open Data (SD-LLOD-17). 26 June.

Gracia, J., Kernerman, I. & Bosque-Gil, J. 2017. Toward linked data-native dictionaries. *Proceedings of ELEX 2017: Lexicography from scratch*. 19-21 September 2017. Available: https://elex.link/elex2017/wp-content/uploads/2017/09/paper33.pdf [2018, January 17].

Gracia, J. & Vila-Suero, D. 2015. *Guidelines for linguistic linked data generation: bilingual dictionaries.* Available: https://www.w3.org/2015/09/bpmlod-reports/bilingual-dictionaries/ [2017, December 25].

Gracia, J., Villegas, M., Gómez-Pérez, A. & Bel, N. 2018. The Apertium bilingual dictionaries on the web of data. *Semantic Web*. 9(2):231-240. Available: http://www.semantic-web-journal.net/system/files/swj1419.pdf [2017, December 31].

Grover, A.S., van Huyssteen, G.B., & Pretorius, M.W. 2011. The South African human language technology audit. *Language Resources and Evaluation.* 45:271-288.

Güldermann, T. 2014. *Towards casting a wider net over N*‖*ng: chances and challenges of archival Khoisan resources.* Available: https://www.iaaw.hu-berlin.de/de/region/afrika/afrika/linguistik/mitarbeiter/1683070/dokumente/2014-03-cape-town-nng-h [2018, June 18].

Hammarström, H., Forkel, R. & Haspelmath, M. 2018. *Glottolog 3.3*. Available: http://glottolog.org/ [2018, June 24].

Hansen, T. & Melnikov, A. 2013. *Additional media type structured syntax suffixes*. Available: https://tools.ietf.org/html/rfc6839 [2017, December 5].

Heath, T. & Bizer, C. 2011. *Linked data: evolving the web into a global data space.* California: Morgan & Claypool Publishers.

Herbert, R.K. & Bailey, R. 2002. The Bantu languages: sociohistorical perspectives. In *Language in South Africa.* R. Mesthrie, Ed. Cambridge: Cambridge University Press. 50-78.

Higgins, S. 2016. Data modelling for analysis, discovery and retrieval. In *Managing digital cultural objects: Analysis, discovery and retrieval.* A. Foster & P. Rafferty, Eds. London: Facet Publishing. 33-34.

Hirst, G. 2009. *Ontology and the lexicon*. Available: ftp://ftp.cs.toronto.edu/pub/gh/Hirst-Ontol-2009.pdf [2018, August 3].

Hirst, G. 2014. Overcoming linguistic barriers to the multilingual semantic web. In *Towards the Multilingual Semantic Web.* P. Buitelaar & P. Cimiano, Eds. Berlin: Springer-Verlag. 3-14.

Hitzler, P., Krötzsch, M. & Rudolph, S. 2010. *Foundations of semantic web technologies.* Boca Raton: Chapman & Hall/CRC.

Hogan, A., Umbrich, J., Harth, A., Cyganiak, R., Polleres, A. & Decker, S. 2012. An empirical survey of linked data conformance. *Web Semantics: Science, Services and Agents on the World Wide Web.* 14:14-44.

How to contribute. n.d. Available: https://wiki.okfn.org/Working_Groups/Linguistics/How_to_contribute [2017, November 2].

Hyland, B. & Villazón Terrazas, B. Eds. 2011. *Cookbook for open government linked data.* Available: https://www.w3.org/2011/gld/wiki/Linked_Data_Cookbook [2018, January 4].

Hyvönen, E. 2012. *Publishing and using cultural heritage linked data on the semantic web.* California: Morgan & Claypool Publishers. DOI: 10.2200/S00452ED1V01Y201210WBE003.

International Organization for Standardization. 1985. *ISO 5807:1985 – Information processing -- Documentation symbols and conventions for data, program and system flowcharts,*

program network charts and system resources charts. Available: https://www.iso.org/standard/11955.html [2018, August 3].

Jurafsky, D. & Martin, J.H. 2009. *Speech and language processing*. 2nd Edition. New Jersey: Pearson Education Inc.

Kay, M. 1997. Multilinguality. In *Survey of the state of the art in human language technology*. G.B. Varile & A. Zampolli, Eds. Cambridge: Cambridge University Press. 245-247.

Keller, M.A., Persons, J., Glaser, H. & Calter, M. 2011. *Report on the Stanford Linked Data Workshop, 27 June – 1 July 2011.* Available: https://www.clir.org/wp-content/uploads/sites/6/LinkedDataWorkshop.pdf [2017, December 26].

Khalfi, M., Nahli, O. & Zarghili, A. 2016. Classical dictionary Al-Qamus in lemon. *Proceedings of 2016 4th IEEE International Colloquium on Information Science and Technology (CiSt).* 24-26 October 2016. DOI:10.1109/CIST.2016.7805065.

Kumar, R. 2014. *Research methodology: a step-by-step guide for beginners.* London: Sage Publications Ltd.

Labra Gayo, J.E., Kontokostas, D. & Auer, S. 2013. Multilingual linked data patterns. *Semantic Web*. 6(4). Available: http://www.semantic-web-journal.net/system/files/swj495.pdf [2017, December 27].

Language: N/u. n.d. Available: http://glottolog.org/resource/languoid/id/nuuu1241 (2018, June 24].

Language: South African Sign Language. n.d. Available: http://glottolog.org/resource/languoid/id/sout1404 (2018, July 1].

Language-subtag-registry. 2018. Available: https://www.iana.org/assignments/language-subtag-registry/language-subtag-registry [2018, June 24].

lemon – The lexicon model for ontologies. n.d. Available: http://lemon-model.net/ [2017, December 18].

León-Araúz, P. & Faber, P. 2014. In *Towards the multilingual semantic web.* P. Buitelaar & P. Cimiano, Eds. Berlin: Springer-Verlag. 31-48.

Levin, M.E. 2014. Language and allergy education: review article. *Current Allergy & Clinical Immunology.* 27(4):290-291.

Lexica: lexica by standards. n.d. Available: http://lod.iula.upf.edu/types/Lexica/by/standards [2018, July 28].

Library of Congress names. n.d. Available: http://id.loc.gov/authorities/names.html [2018, January 10].

Library of Congress subject headings. n.d. Available: http://id.loc.gov/authorities/subjects.html [2018, January 10].

Linguistic linked open data. 2018. Available: http://linguistic-lod.org/ [2018, January 30].

Louw, J.A. 1984. Word categories in Southern Bantu. *African Studies.* 43(1):231-239.

Lowe, D. & Hall, W. 1999. *Hypermedia & the web: an engineering approach.* Chichester: John Wiley & Sons Ltd.

Lucidchart. 2018. *Flowchart symbols and notation.* Available: https://www.lucidchart.com/pages/flowchart-symbols-meaning-explained [2018, August 3].

Mail archive: media types for RDF languages N3 and Turtle. Available: https://www.w3.org/2008/01/rdf-media-types#mail [2017, December 3].

Maxwell, J.A. 1996. *Qualitative research design: an interactive approach.* California: Sage Publications.

Maxwell, J.A. 2013. *Qualitative research design: an interactive approach.* 3rd Edition. California: Sage Publications, Inc.

McArthur, T. 1986. *Worlds of reference*. Cambridge: Cambridge University Press.

McArthur, T. 1998. Guides to tomorrow's English. *English Today.* 14(3):21-26.

McCrae, J. 2012. *LMF.* Available: http://lemon-model.net/lemon-cookbook/node46.html [2017, December 21].

McCrae, J., Aguado-de-Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D. & Wunner, T. 2012. Interchanging lexical resources on the semantic web. *Language Resources & Evaluation.* 46:701-719.

McCrae, J., Montiel-Ponsoda, E. & Cimiano, P. 2012. Integrating WordNet and Wiktionary with lemon. In *Linked Data in Linguistics*. C. Chiarcos, S. Nordhoff & S. Hellman, Eds. Heidelberg: Springer. 25-34.

McCrae, J., Spohr, D. & Cimiano, P. 2011. Linking lexical resources and ontologies on the semantic web with lemon. In *The Semantic Web: Research and Applications.* G. Antoniou, M. Grobelnik, E. Simperl, B. Parsia, D. Plexousakis, P. De Leenheer & J. Pan, Eds. 245-259.

McCrae, J.P., Bosque-Gil, J., Gracia, J., Buitelaar, P. & Cimiano, P. 2017. The Ontolex-Lemon model: development and applications*. Proceedings of ELEX 2017: Lexicography from Scratch*. 19-21 September 2017. Available: https://elex.link/elex2017/wp-content/uploads/2017/09/paper36.pdf [2018, January 6].

McCrae, J.P. & Gracia, J. 2017. Introduction to the Ontolex-Lemon Model [Practical session]. 2nd Summer Datathon on Linguistic Linked Open Data. 26 June.

McCrae, J.P. & Unger, C. 2014. Design patterns for engineering the ontology-lexicon interface. In *Towards the Multilingual Semantic Web.* P. Buitelaar & P. Cimiano, Eds. Berlin: Springer-Verlag. 15-30.

Mesthrie, R. 2002. Introduction. *Language in South Africa.* R. Mesthrie, Ed. Cambridge: Cambridge University Press.

Miles, M.B. & Huberman, A.M. 1994. *Qualitative data analysis.* California: Sage Publications.

Mitchell, E.T. 2013. Building blocks of linked open data in libraries. *Library Technology Reports*. 49(5):11-25.

Montiel-Ponsoda, E., Vila-Suero, D., Villazón-Terrazas, B., Dunsire, G., Escolano Rodríguez, E. & Gómez-Pérez, A. 2011. Style guidelines for naming and labeling ontologies in the multilingual web. *Proceedings of International Conference on Dublin Core and Metadata Applications*. 106-115. Available: http://oa.upm.es/12469/1/INVE_MEM_2011_105132.pdf [2018, January 7].

The Multilingual Morpheme Ontology: home. 2018. Available: http://mmoon.org/ [2018, January 17].

Murata, M., St. Laurent, S., & Kohn, D. 2001. *XML media types*. Available: https://tools.ietf.org/html/rfc3023 [2017, December 5].

National Institute for the Deaf. 2016. *Digital SASL dictionary: thousands to benefit from first SASL thesaurus*. Available: https://www.nid.org.za/digital-sasl-dictionary/ [2018, July 1].

Nazari, M. 2010. Design and process of a contextual study of information literacy: an Eisenhardt approach. *Library & Information Science Research.* 32:179-191.

Ngcobo, M.N. Department of Linguistics and Modern Languages. 2010. *Only study guide for LIN3704: language planning and language description.* Pretoria: University of South Africa.

Njeyiyana, M.S. 2018. South African Sign Language (SASL): evidence of the use of school-lects? [Poster]. International Congress of Linguists. 4 July.

Nkomo, D. 2010. Affirming a role for specialised dictionaries in indigenous African languages. *Lexikos*. 20:371-389.

N|u of N‖ng (ngh). n.d. Available: http://www.multitree.org/codes/ngh-nuu [2018, Jun 20].

Ocholla, D.N. & Le Roux, J. 2011. Conceptions and misconceptions of theoretical frameworks in library and information science research. *Proceedings of the 6th Biennial Prolissa Conference, Pretoria*. 9-11 March 2011. Available: http://www.lis.uzulu.ac.za/2011/Ocholla%20and%20Le%20Roux%20prolissa%20conference%202011%20revised%2016%20March%202011.pdf [2017, December 17].

Olsen, D.R. 2014. *Chinese writing*. Available: https://www.britannica.com/topic/Chinese-writing [2018, July 10].

Onwuegbuzie, A.N. & Frels, R. 2016. *7 Steps to a comprehensive literature review: a multimodal & cultural approach.* California: Sage Publications Ltd.

The Open Archives Initiative protocol for metadata harvesting. 2015. Available: https://www.openarchives.org/OAI/openarchivesprotocol.html [2018, January 3].

Pahl, H.W., Pienaar, A.M. & Ndungane, T.A. Eds. 1989. *The greater dictionary of Xhosa: volume 3.* Alice: University of Fort Hare.

PanSALB. 2015. *PanSALB history*. Available: http://www.pansalb.org/history1.html [2017, November 2].

Phillips, A. & Davis, M., Eds. 2009. *BCP 47: tags for identifying languages.* Available: https://tools.ietf.org/html/bcp47 [2018, June 24].

Phonetic Symbols. 2002. In *Language in South Africa.* R. Mesthrie, Ed. Cambridge: Cambridge University Press. xiii.

Pretorius, L. 2014. The multilingual semantic web as virtual knowledge commons: the case of the under-resourced South African languages. In *Towards the multilingual semantic web.* P. Buitelaar & P. Cimiano, Eds. Berlin: Springer-Verlag. 49-66.

Prévot, L., Huang, C., Calzolari, N., Gangemi, A., Lenci, A. & Oltramari, A. 2010. Ontology and the lexicon: a multidisciplinary perspective. In *Ontology and the Lexicon: A Natural Language Processing Perspective*. 3-24. DOI: 10.1017/CBO9780511676536.002.

Prinsloo, D. 2010. Review: Oxford bilingual school dictionary: Zulu and English. *Lexikos.* 20:760-766.

Prud'hommeaux, E. 2017. *application/n-triples.* Available: https://www.iana.org/assignments/media-types/application/n-triples [2017, December 5].

Rafferty, P. 2016. Managing, searching and finding digital cultural objects: putting it in context. In *Managing digital cultural objects: analysis, discovery and retrieval*. A. Foster & P. Rafferty, Eds. London: Facet Publishing. 3-24.

Raghallaigh, B. & Měchura, M. 2014. Developing high-end reusable tools and resources for Irish-language terminology, lexicography, onomastics (toponymy), folkloristics, and more, using modern web and database technologies. *Proceedings of the First Celtic Language Technology Workshop.* 66-70.

Rasinger, S.M. 2013. Quantitative methods: concepts, frameworks and issues. In *Research methods in linguistics.* L. Litosseliti, Ed. London: Bloomsbury Academic. 49-67.

S00000249n. n.d. Available: http://babelnet.org/rdf/page/s00000249n [2018, August 9].

Sachs, J. & Finin, T. 2010. *What does it mean for a URI to resolve?.* Available: http://ebiquity.umbc.edu/_file_directory_/papers/495.pdf [2017, December 26].

Saint-Dizier, P. 2001. Underspecified lexical conceptual structures for sense variations. *Computing Meaning.* 2:113-128.

Sample OAI-PMH requests. n.d. Available: https://memory.loc.gov/ammem/oamh/oai_request.html [2018, January 3].

Schermer, T. 2016. Lexicon. In *The linguistics of sign languages: an introduction*. A. Baker, B. van den Bogaerde, R. Pfau & T. Schermer, Eds. Amsterdam: John Benjamins Publishing Company. 173-196.

Schoonheim, T. 2014. The European Network of e-Lexicography (ENel). *Kernerman Dictionary News*. 22:1-4.

Shah, S. & Brenzinger, M. 2016. *Ouma Geelmeid ke kx'u ||xa||xa N|uu*. Cape Town: CALDi, Centre for African Language Diversity, University of Cape Town.

Shepherd, R.H.W. 1952. *A South African medical pioneer: the life of Neil MacVicar*. Lovedale: Lovedale Press.

SIL International. 2017. *639 Identifier documentation: ngh*. Available: https://iso639-3.sil.org/code/ngh [2018, June 20].

SIL International. 2017. *639 Identifier documentation: sfs*. Available: https://iso639-3.sil.org/code/sfs [2018, July 1].

SIL International. 2018. *South African Sign Language.* Available: https://www.ethnologue.com/language/sfs [2018, July 1].

Simon, M. 2011. *Assumptions, limitations and delimitations.* Available: http://dissertationrecipes.com/wp-content/uploads/2011/04/AssumptionslimitationsdelimitationsX.pdf [2017, November 5].

Simons, N. & Richardson, J. 2013. *New content in digital repositories: the changing research landscape*. Oxford: Chandos Publishing.

Skyttner, L. 1996. *General systems theory: an introduction.* Hampshire: MacMillan Press Ltd.

Sporny, M., Kellogg, G. & Lanthaler, M. 2013. *application/ld+json*. Available: https://www.iana.org/assignments/media-types/application/ld+json [2017, December 5].

Stake, R.E. 2000. The case study method in social inquiry. In *Case study method.* R. Gomm, M. Hammersley & P. Foster, Eds. London: Sage Publications Ltd.

Statistics South Africa. 2012. *Census 2011 Census in brief.* Pretoria: Statistics South Africa. Available: http://www.statssa.gov.za/census/census_2011/census_products/Census_2011_Census_in_brief.pdf [2017, November 5].

"stomach". *A Dictionary of New Zealand Sign Language.* 1997. Wellington: Bridget Williams Books.

"Stomach". *English-Xhosa dictionary for nurses.* 1935. 2nd ed. Lovedale, South Africa: Lovedale Press.

"stomach". *Oxford English Xhosa dictionary.* 2013. 25th ed. Cape Town: Oxford University Press Southern Africa (Pty) Ltd.

Stuart, D. 2016. *Practical ontologies for information professionals.* London: Facet Publishing.

Subfamily: Nguni (S.40). n.d. Available: http://glottolog.org/resource/languoid/id/ngun1276 [2018, February 11].

Subfamily: West Germanic. n.d. Available: http://glottolog.org/resource/languoid/id/west2793 [2017, December 17].

Taljard, E. & Bosch, S.E. 2006. A comparison of approaches to word class tagging: disjunctively vs. conjunctively written Bantu languages. *Nordic Journal of African Studies.* 15(4):428-442.

Tellis, W. 1997. Introduction to case study. *The Qualitative Report.* 3(2):1-14.

Tennis, J.T. 2007. Scheme versioning in the semantic web. In *Knitting the semantic web.* J. Greenberg & E. Méndez, Eds. New York: The Haworth Information Press. 85-104.

text/turtle. 2011. Available: https://www.iana.org/assignments/media-types/text/turtle [2017, December 5].

Tittel, S. & Chiarcos, C. 2018. Historical lexicography of Old French and linked open data: transforming the resources of the Dictionnaire étymologique de l'ancien francais with Ontolex-Lemon. *Proceedings of the 6th Workshop on Linked Data in Linguistics: Towards Linguistic Data Science, Co-Located with LREC2018, Miyazaki, Japan.* 12 May 2018. Available: http://lrec-conf.org/workshops/lrec2018/W33/pdf/23_W33.pdf [2018, July 28].

Trask, R.L. 1993. *A dictionary of grammatical terms in linguistics.* London: Routledge.

Tshabe, S.L. 2006. *The Greater Dictionary of isiXhosa: volume 1.* Alice: University of Fort Hare.

U.S. National Library of Medicine. 2018. *Medical subject headings.* Available: https://www.nlm.nih.gov/mesh/meshhome.html [2018, November 21].

Van Der Merwe, M. 2015. Giving breath to a dying history. *Daily Maverick.* Available: https://www.dailymaverick.co.za/article/2015-01-23-giving-breath-to-a-dying-history/#.Wyvou9WFMsk [2018, June 21].

Van Erp, M. 2012. Reusing linguistic resources: tasks and goals for a linked data approach. In *Linked Data in Linguistics*. C. Chiarcos, S. Nordhoff & S. Hellman, Eds. Heidelberg: Springer. 57-64.

Van Hooland, S. & Verborgh, R. 2014. *Linked data for libraries, archives and museums*. London: Facet Publishing.

Vila-Suero, D. & Gómez-Pérez, A. 2013. *datos.bne.es and MARiMbA: an insight into Library Linked Data.* Available: http://oa.upm.es/29328/1/INVE_MEM_2013_168001.pdf [2017, December 25].

Vila-Suero, D., Gómez-Pérez, A., Montiel-Ponsoda, E., Gracia, J. & Aguado-de-Cea, G. 2014. Publishing linked data on the web: the multilingual dimension. In *Towards the Multilingual Semantic Web.* P. Buitelaar & P. Cimiano, Eds. Berlin: Springer-Verlag. 101-117.

Villazón-Terrazas, B., Vilches-Blázquez, L.M., Corcho, O. & Gómez-Pérez, A. 2011. *Methodological guidelines for publishing government linked data.* Available: https://www.lri.fr/~hamdi/datalift/tuto_inspire_2012/Suggestedreadings/egovld.pdf [2017, December 25].

Welcome to the New Zealand Sign Language dictionary. n.d. Available: https://nzsl.vuw.ac.nz/ [2018, July 1].

What is the SASL dictionary? n.d. Available: https://nid.org.za/dictionary/ [2018, July 1].

Wood, D., Zaidman, M., Ruth, L. & Hausenblas, M. 2014. *Linked data: structured data on the web*. New York: Manning Publications Co.

Woodside, A.G. & Wilson, E.J. 2003. Case study research methods for theory building. *Journal of Business & Industrial Marketing.* 18(6/7):493-508.

WordNet RDF. n.d. Available: http://wordnet-rdf.princeton.edu/ [2017, November 11].

World Wide Web Consortium. 2004. *Architecture of the world wide web, volume one: W3C recommendation 15 December 2004.* I. Jacobs & N. Walsh, Eds. Available: https://www.w3.org/TR/webarch/ [2018, June 24].

World Wide Web Consortium. 2004. *OWL web ontology language overview: W3C recommendation 10 February 2004*. Available: https://www.w3.org/TR/owl-features/ [2018, January 11].

World Wide Web Consortium. 2008. *Media types issues for test RDF formats.* Available: https://www.w3.org/2008/01/rdf-media-types [2017, December 3].

World Wide Web Consortium. 2009. *SKOS Simple Knowledge Organization System primer: W3C working group note 18 August 2009*. Available: https://www.w3.org/TR/skos-primer/ [2018, January 16].

World Wide Web Consortium. 2013. *PROV-Dictionary: Modeling provenance for dictionary data structures: W3C working group note 30 April 2013*. Available: https://www.w3.org/TR/2013/NOTE-prov-dictionary-20130430/ [2018, January 1].

World Wide Web Consortium. 2013. *PROV-O: The PROV ontology: W3C recommendation 30 April 2013*. Available: https://www.w3.org/TR/prov-o/ [2018, January 1].

World Wide Web Consortium. 2014. *JSON-LD 1.0: W3C recommendation 16 January 2014*. Available: https://www.w3.org/TR/json-ld/ [2017, December 5].

World Wide Web Consortium. 2014. *Language tags in HTML and XML*. Available: https://www.w3.org/International/articles/language-tags/index.en [2018, June 24].

World Wide Web Consortium. 2014. *RDF 1.1 N-Triples: W3C recommendation 25 February 2014.* Available: https://www.w3.org/TR/n-triples/ [2017, December 5].

World Wide Web Consortium. 2014. *RDF 1.1 Turtle: W3C recommendation 25 February 2014*. Available: https://www.w3.org/TR/turtle/ [2017, December 3].

World Wide Web Consortium. 2014. *RDF 1.1 XML syntax: W3C recommendation 25 February 2014*. Available: https://www.w3.org/TR/rdf-syntax-grammar/ [2017, November 11].

World Wide Web Consortium. 2015. *Vocabularies*. Available: https://www.w3.org/standards/semanticweb/ontology [2017, November 5].

Wunner, T. 2012. *LEXINFO vocabulary.* Available: http://vocab.deri.ie/lexinfo# [2018, January 17].

Yin, R.K. 1994. *Case study research: design and methods.* London: Sage Publications.

Yin, R.K. 2014. *Case study research: design and methods.* 5th Edition. California: Sage Publications, Inc.

Zainal, Z. 2007. Case study as a research method. *Jurnal Kemanusiaan.* 9:1-6. Available: http://psyking.net/htmlobj-3837/case_study_as_a_research_method.pdf [2017, November 3].

Zgusta, L. 1971. *Manual of lexicography.* Prague: Academia, Publishing House of the Czechslovak Academy of Sciences.

|*'Auni of N*ǁ*ng (ngh).* n.d. Available: http://www.multitree.org/codes/ngh-nuu [2018, Jun 20].

"|x'â". *Ouma Geelmeid ke kx'u* ||*xa*||*xa N*|*uu.* 2016. Cape Town: CALDi, Centre for African Language Diversity, University of Cape Town.

"ǁx'â". *Ouma Geelmeid ke kx'u* ||*xa*||*xa N*|*uu.* 2016. Cape Town: CALDi, Centre for African Language Diversity, University of Cape Town.

Appendices

Appendix A: Figures

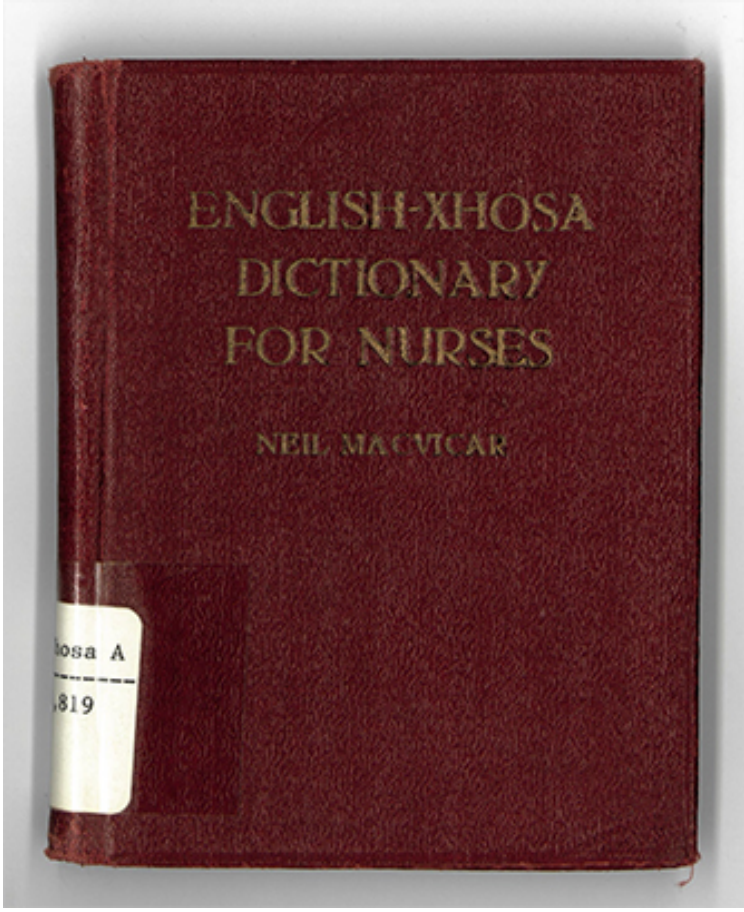
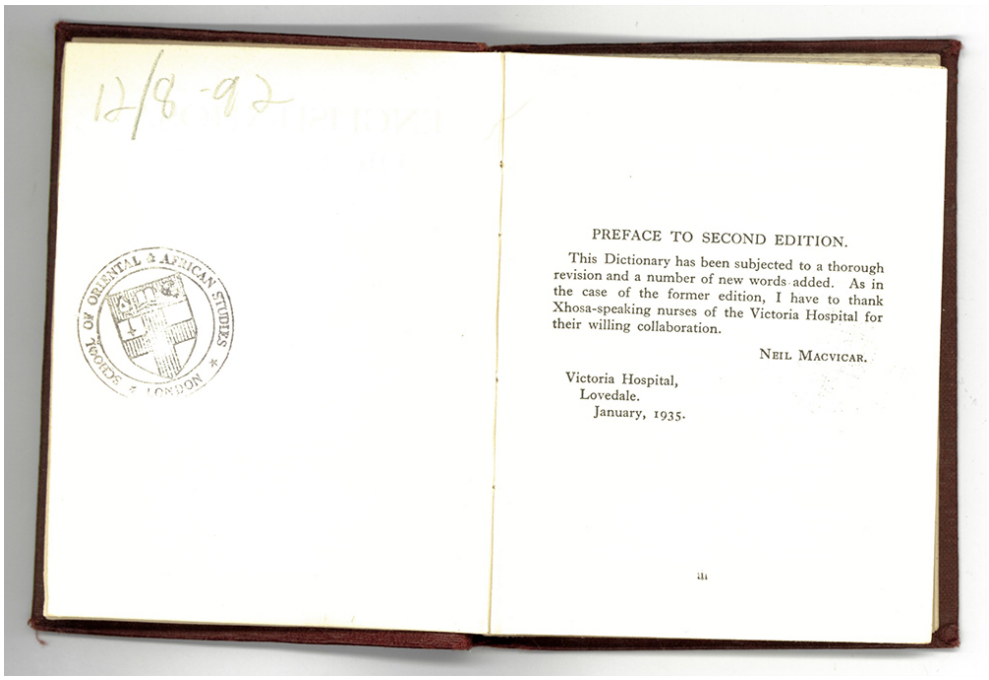


Figure Appendices-1: Scanned image of the book cover of EXDN



PREFACE TO SECOND EDITION.

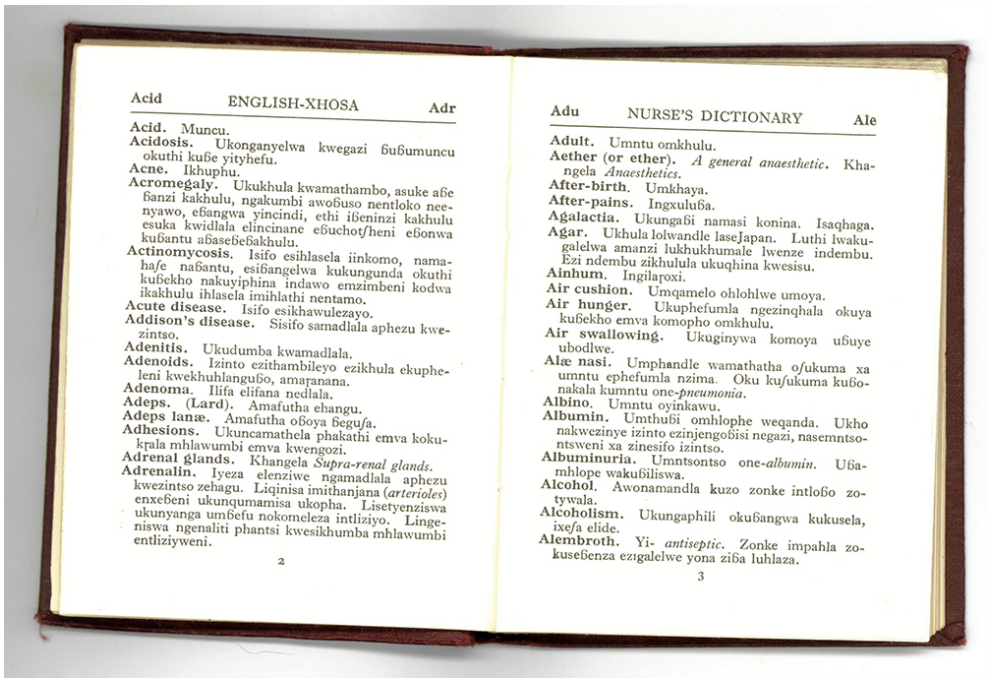
This Dictionary has been subjected to a thorough revision and a number of new words added. As in the case of the former edition, I have to thank Xhosa-speaking nurses of the Victoria Hospital for their willing collaboration.

NEIL MACVICAR.

Victoria Hospital,
Lovedale.
January, 1935.

iii

Figure Appendices-2: Scanned image of the preface (front matter) of EXDN



Acid ENGLISH-XHOSA Adr

Acid. Muncu.
Acidosis. Ukonganyelwa kwegazi бuбumuncu okuthi kuбe yityhefu.
Acne. Ikhuphu.
Acromegaly. Ukukhula kwamathambo, asuke aбe бanzi kakhulu, ngakumbi awoбuso nentloko neenyawo, eбangwa yincindi, ethi iбeninzi kakhulu esuka kwidlala elincinane eбuchot/heni eбonwa kuбantu aбaseбeбakhulu.
Actinomycosis. Isifo esihlasela iinkomo, namaha/e naбantu, esiбangelwa kukungunda okuthi kuбekho nakuyiphina indawo emzimbeni kodwa ikakhulu ihlasela imihlathi nentamo.
Acute disease. Isifo esikhawulezayo.
Addison's disease. Sisifo samadlala aphezu kwezintso.
Adenitis. Ukudumba kwamadlala.
Adenoids. Izinto ezithambileyo ezikhula ekupheleni kwekhuhlanguбo, amaranana.
Adenoma. Ilifa elifana nedlala.
Adeps. (**Lard**). Amafutha ehangu.
Adeps lanæ. Amafutha oбoya бegu/a.
Adhesions. Ukuncamathela phakathi emva kokuкrala mhlawumbi emva kwengozi.
Adrenal glands. Khangela *Supra-renal glands.*
Adrenalin. Iyeza elenziwe ngamadlala aphezu kwezintso zehagu. Liqinisa imithanjana (*arterioles*) enxeбeni ukunqumamisa ukopha. Lisetyenziswa ukunyanga umбefu nokomeleza intliziyo. Lingeniswa ngenaliti phantsi kwesikhumba mhlawumbi entliziyweni.

2

Adu NURSE'S DICTIONARY Ale

Adult. Umntu omkhulu.
Aether (or ether). *A general anaesthetic.* Khangela *Anaesthetics.*
After-birth. Umkhaya.
After-pains. Ingxuluба.
Agalactia. Ukungaбi namasi konina. Isaqhaga.
Agar. Ukhula lolwandle laseJapan. Luthi lwakugalelwa amanzi lukhukhumale lwenze indembu. Ezi ndembu zikhulula ukuqhina kwesisu.
Ainhum. Ingilaroxi.
Air cushion. Umqamelo ohlohlwe umoya.
Air hunger. Ukuphefumla ngezinqhala okuya kuбekho emva komopho omkhulu.
Air swallowing. Ukuginywa komoya uбuye ubodlwe.
Alæ nasi. Umphandle wamathatha o/ukuma xa umntu ephefumla nzima. Oku ku/ukuma kuбonakala kumntu one-*pneumonia.*
Albino. Umntu oyinkawu.
Albumin. Umthuбi omhlophe weqanda. Ukho nakwezinye izinto ezinjengoбisi negazi, nasemntsontsweni xa zinesifo izintso.
Albuminuria. Umntsontso one-*albumin.* Uбamhlope wakuбiliswa.
Alcohol. Awonamandla kuzo zonke intloбo zotywala.
Alcoholism. Ukungaphili okuбangwa kukusela, ixe/a elide.
Alembroth. Yi- *antiseptic.* Zonke impahla zokuseбenza ezigalelwe yona ziba luhlaza.

3

Figure Appendices-3: Scanned image of an example of the central list

Appendix B: Serialisation formats for RDF

The namespace `londi` (a prefix for URIs beginning with "//londisizwe.org/") is defined for the case study (W3C, 2014e). Using content negotiation, a machine-readable and human-readable view is presented for each URI associated with the `londi` namespace. However, due to time constraints of the study, a human-friendly RDF serialisation was selected for the machine-readable view, and the human-readable view defaults to the machine-readable view. For the RDF dataset, which will be uploaded to several data centres (see Section 1.6), a machine-friendly serialisation was selected. For this reason, serialisations which are intended for character display only, namely those whose MIME types begin with "text/", are not considered for the RDF dataset (W3C, 2008; "Mail archive: media …", 2008).

Of the four formats presented below, the Turtle serialisation was handwritten and the other formats were automatically generated from the Turtle format, using http://www.easyrdf.org. In terms of human-readability, both RDF/XML and JSON-LD are complex to interpret; Turtle is the most readable, followed by the N-Triples format (Heath & Bizer, 2011:18-19). For the Turtle format, the triples are serialised as short sentences, with each triple terminated by a ".", and prefixes used for the namespaces (W3C, 2014d). If there is more than one triple for the same subject, the subject can be omitted, and the predicate and object of each triple can follow, each separated by a ";" (W3C, 2014d). As Turtle is human-friendly – both for writing the RDF by hand and for its readability, it has been selected as the format for the dereferenceable URIs (Hyvönen, 2012:22; Heath & Bizer, 2011:19).

As a human-readable format, the N-Triples format is still readable, although prefixes cannot be used to represent namespaces. Each triple is written one per line, with each URI in full, and the triple is terminated by a "." (W3C, 2014c). BabelNet, the multilingual encyclopaedic dictionary (with over 1.1 billion triples by the year 2014), serialises their RDF in N-Triples, declaring it to be the "best for huge data sets" ("Converting BabelNet as …", 2014; Ehrmann et al., 2014:406), a view shared by Heath and Bizer (2011:20). When loading data into a triplestore, with this notation, it can be parsed one triple at a time without requiring the whole dataset to be loaded into memory

(Hyvönen, 2012:22; Heath & Bizer, 2011:20). The N-Triples format has thus been selected as the format for the RDF dataset.

Turtle

For this format, the MIME type is "text/turtle", the file extension is ".ttl", and the character encoding is UTF-8 (W3C, 2014d; "text/turtle", 2011). This is a format for the human user, intended for display purposes ("text/turtle", 2011; Hyvönen, 2012:22).

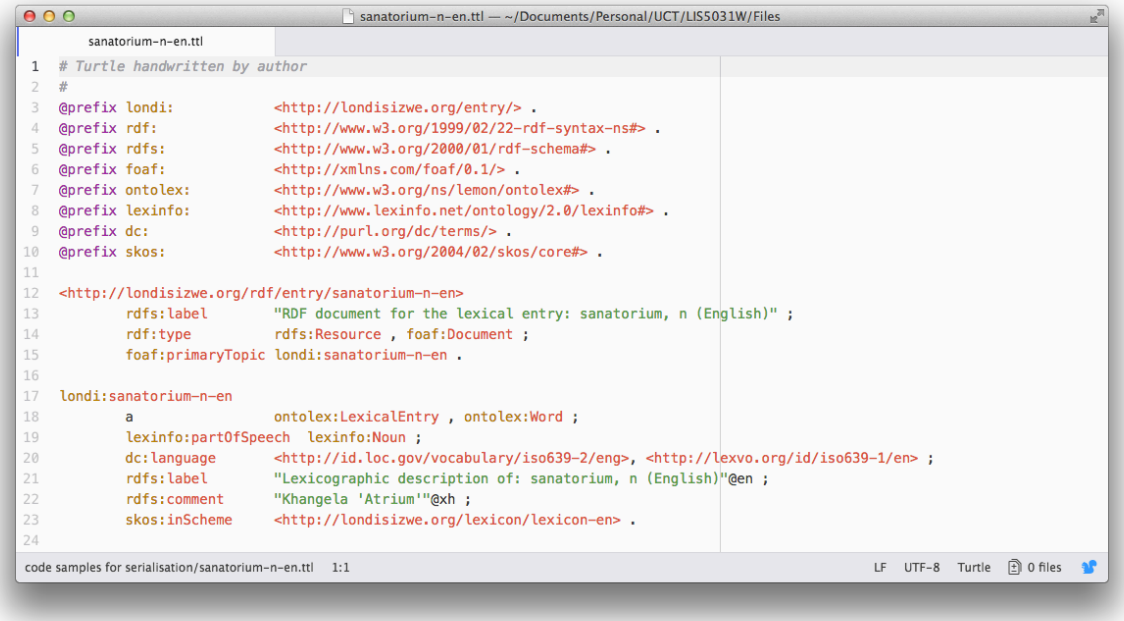


Figure Appendices-4: Code sample for Turtle RDF syntax

RDF/XML

The MIME type is "application/RDF+XML" and the file extension is ".rdf" (W3C, 2014e). This format is intended for machine-processing, for use by web user agents (for example, browsers and crawlers) and other applications ("application/rdf+xml", 2004). The recommended character encoding is UTF-8 and UTF-16 (Murata, St. Laurent & Kohn, 2001:9).

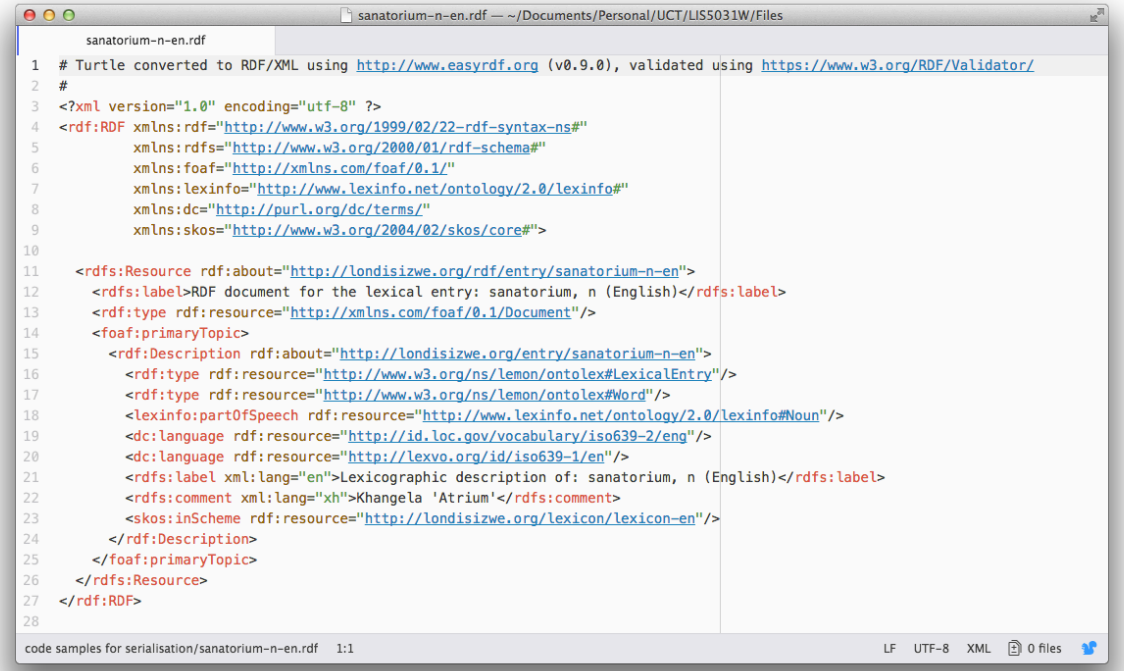


Figure Appendices-5: Code sample for RDF/XML RDF syntax

JSON-LD

The MIME type is "application/ld+json" and the file extension is ".jsonld" (W3C, 2014a; Sporny, Kellogg & Lanthaler, 2013). This format is intended for machine-processing, for use by applications written in languages such as JavaScript, Python, Ruby, and PHP (Sporny, Kellogg & Lanthaler, 2013). The character encoding is UTF-8, UTF-16, or UTF-32 (Sporny, Kellogg & Lanthaler, 2013; Hansen & Melnikov, 2013:3).

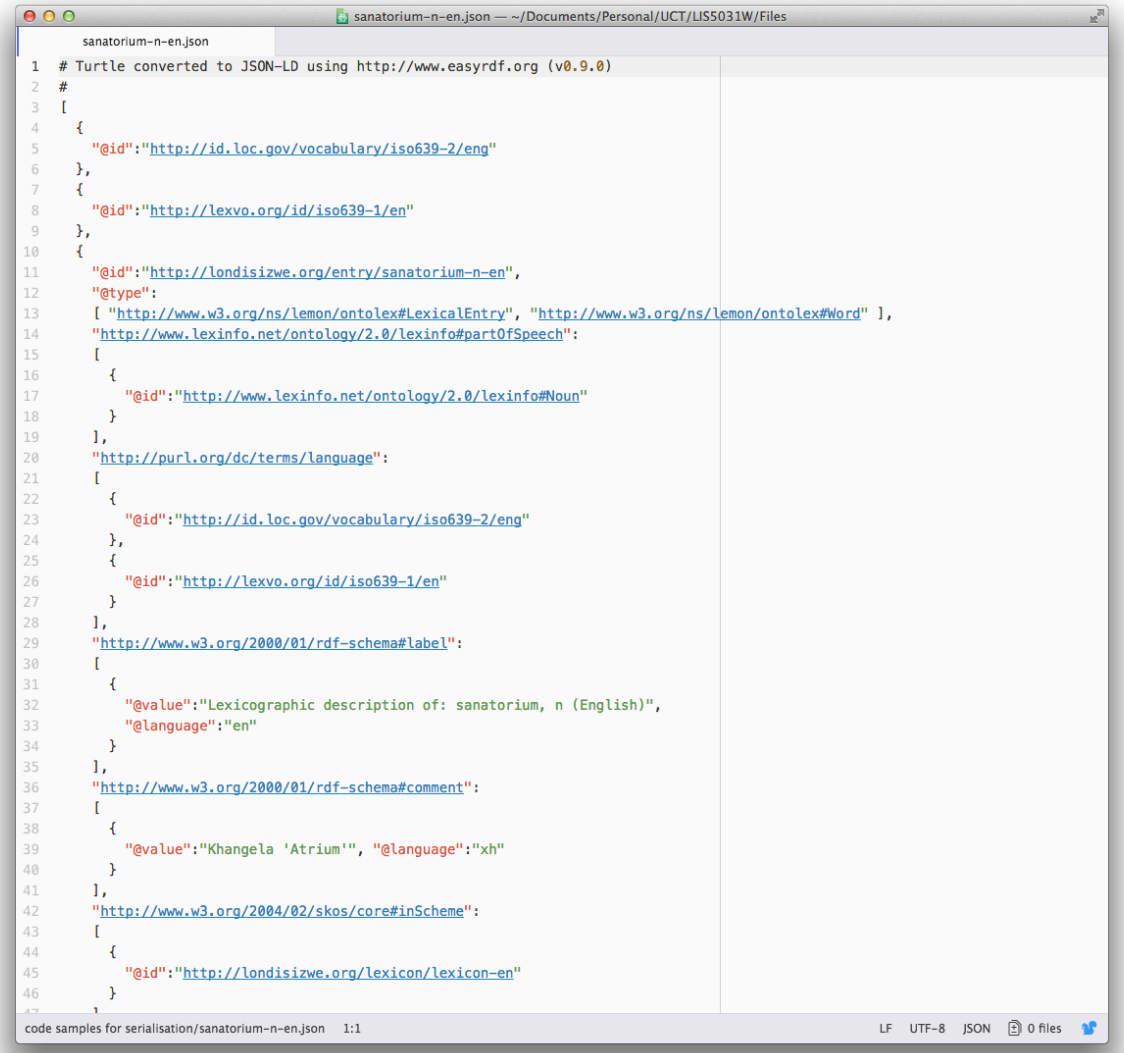


Figure Appendices-6: Code sample for JSON-LD RDF syntax

N-Triples

The MIME type is "application/n-triples", the file extension is ".nt", and the character encoding is UTF-8 (W3C, 2014c; Prud'hommeaux, 2017). This format is intended for both human consumption and machine-processing, and it is the *de facto* standard when working with large datasets (W3C, 2014c; Heath & Bizer, 2011:20).

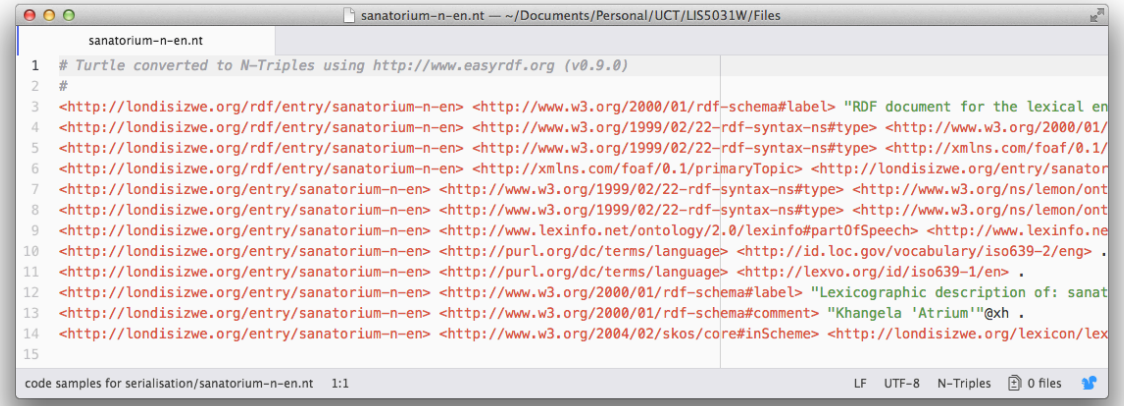


Figure Appendices-7: Code sample for N-Triples RDF syntax

Appendix C: Flowchart modelling

The symbols used in flowchart modelling are defined in ISO 5807:1985 (International Organization for Standardization, 1985), however because this standard is behind a paywall, the symbols as defined by Lucidchart have been used.

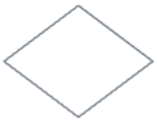
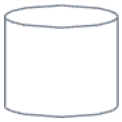
	Description
	*Terminator symbol* This symbol indicates the start and end of a flow (Lucidchart, 2018).
	*Process symbol* This symbol represents "a process, action, or function" (Lucidchart, 2018).
	*Decision symbol* This symbol indicates a question, with yes and no leading to different branches of the diagram (Lucidchart, 2018).
	*Manual input symbol* This symbol represents manual input by the user (Lucidchart, 2018).
	*Database symbol* This symbol represents a database or storage service (Lucidchart, 2018).
	*Document symbol* This symbol represents a document, and it can be an input or an output (Lucidchart, 2018).
	*Multiple documents symbol* This symbol represents multiple documents (Lucidchart, 2018).

Table Appendices-1: Symbols used in flowchart modelling

Appendix D: isiXhosa Noun Classes

The noun classes are derived from a table provided in Volume 1 of The Greater Dictionary of IsiXhosa (Tshabe, 2006:xiv).

Class Number	Prefix	Examples of Nouns	Type
1	um-	umntu (a person), umongi (a nurse)	Singular
1(a)	u-	umama (my mother), uKuhle (Kuhle)	Singular
2	aba-, abe-, ab-	abantu (people), abongi (nurses)	Plural, of Cl 1
2(a)	oo-	oomama (mother, my mother and company), ooKuhle (Kuhle and company)	Plural, of Cl 1(a)
3	um-	umzi (a home, homestead), umbhalo (writing a document)	Singular
4	imi-	imizi, imibhalo	Plural, of Cl 3
5	ili-, i-	ilizwi (a voice, word), ilihlo (an eye)	Singular
6	ama-, ame-	amazwi, amehlo	Plural, of Cl 5
7	isi-, is-, isa-	isitya (a bowl), isono (a sin), isazela (conscience, feeling of guilt)	Singular
8	izi-, iz-, iza-	izitya, izono, izazela	Plural, of Cl 7
9	i-	inja (a dog), imvubo (hippopotamus)	Singular
10	izi-, ii-	izinja (dogs), iimvubu (hippotami), izimvo (opinions)	Plural of Cl 9, 11, and sometimes, Cl 14
11	ulu-, u-	uluvo (opinion), uthando (love)	Singular. Its plural is in Cl 10, although some nouns are abstract
12	*	-	-
13	*	-	-
14	ubu-, ub-, u-	ubongikazi (nursing profession; state of being a nurse), ubom / ubomi (life)	Singular. Some nouns are abstract; most do not have plural forms
15	uku-, ukw-, uk-	ukutya, ukwanda, ukwindla, ukona	Singular, infinitive form.

Table Appendices-2: isiXhosa noun classes

Appendix E: Londisizwe Noun Class Vocabulary

```
@prefix :     <https://ontology.londisizwe.org/nounclass#> .
@prefix owl:  <http://www.w3.org/2002/07/owl#> .
@prefix rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix xml:  <http://www.w3.org/XML/1998/namespace> .
@prefix xsd:  <http://www.w3.org/2001/XMLSchema#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

<https://ontology.londisizwe.org/nounclass> rdf:type owl:Ontology .

#################################################################################################################
#    Object Properties
#################################################################################################################

:inNounClass
    rdf:type        owl:ObjectProperty ;
    rdfs:range      :IsiXhosaNounClass ;
    rdfs:label      "inNounClass"@en .

#################################################################################################################
#    Classes
#################################################################################################################

:IsiXhosa
    rdf:type        owl:Class ;
    rdfs:label      "IsiXhosa"@en .

:IsiXhosaNounClass
    rdf:type        owl:Class ;
    rdfs:subClassOf :IsiXhosa ;
    rdfs:label      "IsiXhosa Noun Class"@en .

:IsiXhosaNC1
    rdf:type        owl:Class ;
    rdfs:subClassOf :IsiXhosaNounClass ;
    rdfs:comment    "The isiXhosa Noun Class 1 is singular. An example is umntu (a person), umongi (a
                    nurse). The prefix is um-"^^xsd:string ;
    rdfs:label      "IsiXhosa NC 1"@en .

:IsiXhosaNC1a
    rdf:type        owl:Class ;
    rdfs:subClassOf :IsiXhosaNounClass ;
    rdfs:comment    "The isiXhosa Noun Class 1(a) is singular. An example is umama (my mother), uKuhle
                    (Kuhle). The prefix is u-"^^xsd:string ;
    rdfs:label      "IsiXhosa NC 1(a)"@en .

:IsiXhosaNC2
    rdf:type        owl:Class ;
    rdfs:subClassOf :IsiXhosaNounClass ;
    rdfs:comment    "The isiXhosa Noun Class 2 is plural of Noun Class 1. An example is abantu
                    (people), abongi (nurses). The prefix is aba-, abe-, ab-"^^xsd:string ;
    rdfs:label       "IsiXhosa NC 2"@en .

:IsiXhosaNC2a
    rdf:type        owl:Class ;
    rdfs:subClassOf :IsiXhosaNounClass ;
    rdfs:comment    "The isiXhosa Noun Class 2(a) is plural of Noun Class 1(a). An example is oomama
                    (mother, my mother and company), ooKuhle (Kuhle and company). The prefix is
                    oo-"^^xsd:string ;
    rdfs:label "IsiXhosa NC 2(a)"@en .

:IsiXhosaNC3
    rdf:type        owl:Class ;
    rdfs:subClassOf :IsiXhosaNounClass ;
```

```
    rdfs:comment    "The isiXhosa Noun Class 3 is singular. An example is umzi (a home, homestead),
                    umbhalo (writing a document). The prefix is um-"^^xsd:string ;
    rdfs:label      "IsiXhosa NC 3"@en .

:IsiXhosaNC4
    rdf:type        owl:Class ;
    rdfs:subClassOf :IsiXhosaNounClass ;
    rdfs:comment    "The isiXhosa Noun Class 4 is plural of Noun Class 3. An example is imizi,
                    imibhalo. The prefix is imi-"^^xsd:string ;
    rdfs:label      "IsiXhosa NC 4"@en .

:IsiXhosaNC5
    rdf:type        owl:Class ;
    rdfs:subClassOf :IsiXhosaNounClass ;
    rdfs:comment    "The isiXhosa Noun Class 5 is singular. An example is ilizwi (a voice, word),
                    ilihlo (an eye). The prefix is ili-, i-"^^xsd:string ;
    rdfs:label      "IsiXhosa NC 5"@en .

:IsiXhosaNC6
    rdf:type        owl:Class ;
    rdfs:subClassOf :IsiXhosaNounClass ;
    rdfs:comment    "The isiXhosa Noun Class 6 is plural of Noun Class 5. An example is amazwi,
                    amehlo. The prefix is ama-, ame-"^^xsd:string ;
    rdfs:label      "IsiXhosa NC 6"@en .

:IsiXhosaNC7
    rdf:type        owl:Class ;
    rdfs:subClassOf :IsiXhosaNounClass ;
    rdfs:comment    "The isiXhosa Noun Class 7 is singular. An example is isitya (a bowl), isono (a
                    sin), isazela (conscience, feeling of guilt). The prefix is isi-, is-, isa-
                    "^^xsd:string ;
    rdfs:label      "IsiXhosa NC 7"@en .

:IsiXhosaNC8
    rdf:type        owl:Class ;
    rdfs:subClassOf :IsiXhosaNounClass ;
    rdfs:comment    "The isiXhosa Noun Class 8 is plural of Noun Class 7. An example is izitya, izono,
                    izazela. The prefix is izi-, iz-, iza-"^^xsd:string ;
    rdfs:label      "IsiXhosa NC 8"@en .

:IsiXhosaNC9
    rdf:type        owl:Class ;
    rdfs:subClassOf :IsiXhosaNounClass ;
    rdfs:comment    "The isiXhosa Noun Class 9 is singular. An example is inja (a dog), imvubo
                    (hippopotamus). The prefix is i-"^^xsd:string ;
    rdfs:label      "IsiXhosa NC 9"@en .

:IsiXhosaNC10
    rdf:type        owl:Class ;
    rdfs:subClassOf :IsiXhosaNounClass ;
    rdfs:comment    "The isiXhosa Noun Class 10 is plural of Noun Class 9, Noun Class 11 and sometimes
                    Noun Class 14. An example is izinja (dogs), iimvubu (hippotami), izimvo
                    (opinions). The prefix is izi-, ii-"^^xsd:string ;
    rdfs:label      "IsiXhosa NC 10"@en .

:IsiXhosaNC11
    rdf:type        owl:Class ;
    rdfs:subClassOf :IsiXhosaNounClass ;
    rdfs:comment    "The isiXhosa Noun Class 11 is singular. Its plural is in Noun Class 10, although
                    some nouns are abstract. An example is uluvo (opinion), uthando (love). The prefix
                    is ulu-, u-"^^xsd:string ;
    rdfs:label      "IsiXhosa NC 11"@en .

:IsiXhosaNC14
    rdf:type        owl:Class ;
    rdfs:subClassOf :IsiXhosaNounClass ;
    rdfs:comment    "The isiXhosa Noun Class 14 is singular. An example is ubongikazi (nursing
                    profession; state of being a nurse), ubom / ubomi (life). The prefix is ubu-, ub-,
```

```
                    u-"^^xsd:string ;
    rdfs:label      "IsiXhosa NC 14"@en .

:IsiXhosaNC15
    rdf:type        owl:Class ;
    rdfs:subClassOf :IsiXhosaNounClass ;
    rdfs:comment    "The isiXhosa Noun Class 15 is singular, of infinitive form. An example is ukutya,
                    ukwanda, ukwindla, ukona. The prefix is uku-, ukw-, uk-"^^xsd:string ;
    rdfs:label      "IsiXhosa NC 15"@en .
```

Appendix F: Clicks

	1 Dental	**2** Palatal	**3** Alveolar	**4** Lateral
isiXhosa	c	q		x
Khoisan (for eg. N|uu)	|	ǂ	!	‖

Table Appendices-3: The clicks of isiXhosa and Khoisan ("Phonetic symbols", 2002:xiii)