

LREC 2020 Workshop
Language Resources and Evaluation Conference
11–16 May 2020

**Workshop on Multimodal Wordnets
(MMWN-2020)**

PROCEEDINGS

Thierry Declerck, Itziar Gonzalez-Dios, German Rigau (eds)

**Proceedings of the LREC 2020
Workshop on Multimodal Wordnets
(MMWN-2020)**

Edited by: Thierry Declerck, Itziar Gonzalez-Dios, German Rigau

ISBN: 979-10-95546-41-2

EAN: 9791095546412

For more information:

European Language Resources Association (ELRA)

9 rue des Cordelières

75013, Paris

France

<http://www.elra.info>

Email: lrec@elda.org

© European Language Resources Association (ELRA)

These workshop proceedings are licensed under a Creative Commons
Attribution-NonCommercial 4.0 International License

Preface

The Global WordNet Association (GWA) is a free, public and non-commercial organization that provides a platform for discussing, sharing and connecting wordnets for all languages in the world. GWA created in 2019 a Working Group (WG) dedicated to multimodal wordnets in order to extend the development and use of wordnets to modalities beyond text. As an initiative of this WG, the Multimodal Wordnets workshop was organised as a satellite event to the twelfth edition of the International Conference on Language Resources and Evaluation (LREC) 2020.

The main objective of this half-day workshop was to initiate the study of the interaction and cross-fertilization between wordnets and existing multimodal resources.

Unfortunately, due to the global issue of the COVID-19 outbreak, the LREC event and the associated workshops could not be held in May in Marseille. However, the Multimodal Wordnets workshop proceedings are published in order to acknowledge and present all the work done by the authors and reviewers.

Seven papers were accepted about the following topics:

- Itziar Gonzalez-Dios, Javier Alvez and German Rigau describe an approach for an ontological organization of qualities based on WordNet adjectives for SUMO-based ontologies.
- Soumya Mohapatra, Shikhar Agnihotri, Apar Garg, Praveen Shah and Shampa Chakraverty present an extension of IndoWordnet with dialectal variants and information.
- John Philip McCrae, Alexandre Rademaker, Ewa Rudnicka and Francis Bond introduce the development under an open-source paradigm of a new version of the English WordNet.
- Jon Alkorta and Itziar Gonzalez-Dios propose a novel approach towards enriching the adjective category in the Basque WordNet by means of a sentiment lexicon.
- Jacek Marciniak presents a solution for a multimodal data organisation based on the the wordnet structure.
- Alexandre Tessarollo and Alexandre Rademaker describe the extension of English WordNet with lithological information based on an authoritative thesaurus.
- Thierry Declerck reports on current work on adding pronunciation information to Wordnets by experimenting with Odenet (Open German WordNet) and Wiktionary.

We hope that the content of those papers can be presented in the context of a future LREC or of relevant events.

Thierry Declerck, Itziar Gonzalez-Dios, German Rigau
Saarbrücken and Donostia, May 2020

Organizers:

Itziar Gonzalez-Dios, University of the Basque Country UPV/EHU
Francis Bond, Nanyang Technological University
Thierry Declerk, German Research Center for Artificial Intelligence
Christiane Fellbaum, Princeton University
Alexandre Rademaker, IBM Research Brazil and EMap/FGV
German Rigau, University of the Basque Country UPV/EHU
Piek Vossen, VU University Amsterdam

Program Committee:

Manex Agirrezabal, University of Copenhagen
Izaskun Aldezabal, University of the Basque Country UPV/EHU
Gorka Azkune, University of the Basque Country UPV/EHU
Sonja Bosch, Department of African Languages, University of South Africa
Federico Boschetti, Institute of computational linguistics "Antonio Zampolli" (ILC-CNR)
Luis Chiruzzo, Universidad de la República de Uruguay
Francesca Frontini, Université Paul Valéry, Montpellier
Xavier Gómez Guinovart, Universidade de Vigo
Fahad Khan, ILC-CNR, Italy
Maria Koutsombogera, Trinity College Dublin
Robert Krovetz, Lexical Research
John P. McCrae, National University of Ireland
Gerard de Melo, Rutgers University
Verginica Mititelu, Romanian Academy Research Institute for Artificial Intelligence
Terhi Nurmikko-Fuller, Australian National University
Ahti Lohk, Tallinn University of Technology
Petya Osenova, Sofia University
Patrizia Paggio, Copenhagen University
Bolette Pedersen, University of Copenhagen
Maciej Piasecki, Wrocław University of Technology
Eli Pociello, Elhuyar Fundazioa
Ronald Poppe, Utrecht University
Andrea Amelio Ravelli, University of Florence
Kevin Scannell, Saint Louis University
Aitor Soroa, University of the Basque Country UPV/EHU

Table of Contents

<i>Towards modelling SUMO attributes through WordNet adjectives: a Case Study on Qualities</i> Itziar Gonzalez-Dios, Javier Alvez and German Rigau	1
<i>Incorporating Localised Context in Wordnet for Indic Languages</i> Soumya Mohapatra, Shikhar Agnihotri, Apar Garg, Praveen Shah and Shampa Chakraverty	7
<i>English WordNet 2020: Improving and Extending a WordNet for English using an Open-Source Methodology</i> John Philip McCrae, Alexandre Rademaker, Ewa Rudnicka and Francis Bond	14
<i>Exploring the Enrichment of Basque WordNet with a Sentiment Lexicon</i> Itziar Gonzalez-Dios and Jon Alkorta	20
<i>Wordnet As a Backbone of Domain and Application Conceptualizations in Systems with Multimodal Data</i> Jacek Marciniak	25
<i>Inclusion of Lithological terms (rocks and minerals) in The Open Wordnet for English</i> Alexandre Tessarollo and Alexandre Rademaker	33
<i>Adding Pronunciation Information to Wordnets</i> Thierry Declerck, Lenka Bajcetic and Melanie Siegel.....	39

Workshop Program
As it was planned, but due to the Covid-19 outbreak
the LREC conference and the associated workshops
could not take place in May 2020 in Marseilles

Towards modelling SUMO attributes through WordNet adjectives: a Case Study on Qualities

Itziar Gonzalez-Dios, Javier Alvez and German Rigau

Incorporating Localised Context in Wordnet for Indic Languages

Soumya Mohapatra, Shikhar Agnihotri, Apar Garg, Praveen Shah and Shampa Chakraverty

English WordNet 2020: Improving and Extending a WordNet for English using an Open-Source Methodology

John Philip McCrae, Alexandre Rademaker, Ewa Rudnicka and Francis Bond

Exploring the Enrichment of Basque WordNet with a Sentiment Lexicon

Itziar Gonzalez-Dios and Jon Alkorta

Wordnet As a Backbone of Domain and Application Conceptualizations in Systems with Multimodal Data

Jacek Marciniak

Inclusion of Lithological terms (rocks and minerals) in The Open Wordnet for English

Alexandre Tessarollo and Alexandre Rademaker

Adding Pronunciation Information to Wordnets

Thierry Declerck, Lenka Bajcetic and Melanie Siegel

Towards modelling SUMO attributes through WordNet adjectives: a Case Study on Qualities

Itziar Gonzalez-Dios, Javier Álvez, German Rigau
 Ixa Group – HiTZ Center, LoRea Group, Ixa Group – HiTZ Center
 University of the Basque Country UPV/EHU
 {itziar.gonzalezd, javier.alvez, german.rigau}@ehu.eus

Abstract

Previous studies have shown that the knowledge about attributes and properties in the SUMO ontology and its mapping to WordNet adjectives lacks of an accurate and complete characterization. A proper characterization of this type of knowledge is required to perform formal commonsense reasoning based on the SUMO properties, for instance to distinguish one concept from another based on their properties. In this context, we propose a new semi-automatic approach to model the knowledge about properties and attributes in SUMO by exploiting the information encoded in WordNet adjectives and its mapping to SUMO. To that end, we considered clusters of semantically related groups of WordNet adjectival and nominal synsets. Based on these clusters, we propose a new semi-automatic model for SUMO attributes and their mapping to WordNet, which also includes polarity information. In this paper, as an exploratory approach, we focus on qualities.

Keywords: Adjectives, WordNet, SUMO, Commonsense Reasoning

1. Introduction

Adjectives are words that express qualities and properties and usually modify nouns. They have been usually studied from a syntactic and lexico-semantic point of view. In WordNet (Fellbaum, 1998) adjectives are derived into two classes: descriptive and relational. Descriptive adjectives establish to their related head nouns values of (typically) bipolar attributes and consequently are organized in terms of binary oppositions (antonymy) and similarity of meaning (synonymy). For instance, the synsets hot_a^1 and $cold_a^1$ are related by the semantic relation *antonym* in WordNet¹. Moreover, each of these adjectives is linked to semantically similar adjectives by similarity. These comparable adjectives are called satellites. In Figure 1, we present the bipolar adjective cluster structure formed by hot_a^1 and $cold_a^1$ and their respective satellites.

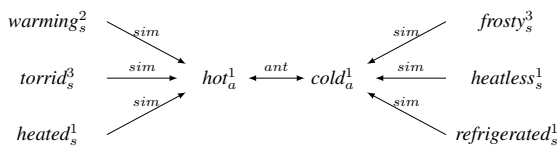


Figure 1: Example of a bipolar adjective cluster

Thus, in a commonsense reasoning scenario, descriptive adjectives need to be represented as attributes of certain nominal and verbal concepts. Therefore, it is necessary to study where this type of adjectives can be used as attributes or properties. Following the example of the pair hot_a^1 and $cold_a^1$, this means that they are possible values of *temperature*. Previous studies have shown that the knowledge about attributes and properties in the SUMO ontology (Niles and Pease, 2001) and its mapping to WordNet adjectives (Niles

and Pease, 2003) lacks of an accurate and complete characterization (Álvez et al., 2019a). For instance, many WordNet adjectives have been mapped to SUMO processes instead to SUMO attributes. A proper characterization of this type of knowledge is required to perform formal commonsense reasoning based on the attributes encoded in SUMO, for example, if we want to distinguish one concept from another based on their properties.

In this framework, two main problems arise when reasoning with the SUMO knowledge related to WordNet adjectives and their antonymy relations. The first one is related to the SUMO mapping and the second one is related to an incomplete axiomatization.

Regarding the mapping, antonymous synset pairs such as $certain_a^3$ and $uncertain_a^2$ are mapped to the same SUMO concept, in this case, to the predicate *knows*. As they are under the same SUMO concept and no contrariness is stated, it is not possible to infer the attributes they express are opposite to each other.

Concerning the under-specification, antonym synset pairs such as $beautiful_a^1$ and $ugly_a^1$ are mapped to the SUMO classes of attributes *SubjectiveStrongPositiveAttribute* and *SubjectiveStrongNegativeAttribute* respectively. Looking at the name of the labels, it seems that the contrariness is expressed, but the only information relating these classes in the ontology is that they are subclasses of *SubjectiveAssessmentAttribute*. Therefore, the ontology is under axiomatized regarding the contrary attribute information.

In this work, we present a case study on qualities and their related adjectives with the aim of improving SUMO and their mapping to WordNet. To that end, we construct adjectival-nominal clusters from WordNet and based on these clusters we create new semantic relations in the Multilingual Central Repository (MCR) (Gonzalez-Agirre et al., 2012) and classes in the Adimen-SUMO ontology (Álvez et al., 2012).

The contributions of this exploratory paper are: a) a de-

¹In this paper, we will refer to the synsets using the format $word_p^s$, where s is the sense number and p is the part-of-speech: n for nouns and a for adjectives.

tailed analysis of adjectival clusters of qualities b) new etymology and morphology based relations for wordnets with the aim of making explicit to which concept attributes should be applied, c) an axiomatization model for qualities and d) a mapping proposal that includes polarity information.

This paper is structured as follows: in Section 2, we present the works related to adjectives in wordnets and in ontologies; in Section 3 we present our knowledge framework; in Section 4 we introduce the improvements proposed for the knowledge about adjectives; in Section 5 we validate our new proposal and, finally, in Section 6 we conclude and outline the future work.

2. Related Work

In this section we provide a brief overview of the approaches used in different lexical knowledge bases and ontologies for representing and exploiting adjectives. The adjectives in the English WordNet (Fellbaum, 1998) are divided into descriptive and relational adjectives. The basic relation between descriptive adjectives is antonymy (direct or indirect). Moreover, by similarity they are linked to semantically comparable adjectives, which are called satellites. This way, bipolar cluster are formed as the one presented in Figure 1. Relational adjectives are also related to nouns and color adjectives are regarded as a special case (Fellbaum et al., 1993). Furthermore, in the morphosemantic links (Fellbaum et al., 2007) adjectives are related to their derived/derivative nouns and verbs. In GermaNet (Hamp and Feldweg, 1997; Henrich and Hinrichs, 2010) the cluster-approach is not followed: adjectives are hierarchically structured, as in the case of nouns and verbs, and, thus, the relation of indirect antonyms is eliminated. Moreover, adjectives are categorised into different semantic classes such as perceptual, spatial, or weather-related.² Building on GermaNet adjectival classification, Tsvetkov et al. (2014) propose supersense (high-level semantic classes) taxonomy for English adjectives. They distinguish 11 classes such as motion, substance or weather³. Regarding the ontologies, the SIMPLE Ontology (Peters and Peters, 2000) distinguishes the adjectives according to their predicative function: intensional adjectives and extensional adjectives. Intensional adjectives have the following subclasses: temporal, modal, emotive, manner, object-related, and emphasisizer. The subclasses of the extensional adjectives are: psychological property, social property, physical property, temporal property, intensifying property, and relational property. The DOLCE family of ontologies relates qualities as individuals to regions, that belong to quality spaces (Gangemi et al., 2016) e.g. *hasQuality(AmazonRiver,wide)*. The Suggested Upper Merged Ontology (SUMO) (Niles and Pease, 2001) and, therefore, its First-order logic conversion Adimen-SUMO (Álvarez et al., 2012) has a class called *Attribute* that includes all qualities, properties, etc. As SUMO is linked to WordNet (Niles

²The full classification can be found in this link: <http://www.sfs.uni-tuebingen.de/GermaNet/adjectives.shtml#Adjective%20Classes>.

³Data can be found in this link: www.cs.cmu.edu/~ytsvetko/adj-supersenses.gz

and Pease, 2003), the adjectives in WordNet fall into *Attribute* and its subclasses such as *SubjectiveAssessmentAttribute*, *SubjectiveStrongNegativeAttribute*, *ShapeAttribute*, or *SubjectiveWeakPositiveAttribute*. In the ontology, these classes are poorly axiomatized and, therefore, we can consider them as *underspecified*.

With respect to its exploitation, the knowledge related to adjectives in WordNet and its mapping into SUMO have been used for semi-automatically creating a large commonsense reasoning benchmark for SUMO-based ontologies (Álvarez et al., 2019b). For this purpose, the authors base on the relations about adjectives *antonymy* and *similarity*, and also considered other relations such as *hyponymy*, which relates noun synsets. Álvarez et al. (2019a) perform a detailed analysis of the experimental results obtained using the proposed benchmark with the objective of shedding light on the commonsense reasoning capabilities of both the benchmark and the involved knowledge resources. One the main reported conclusions is that among the analyzed problems only 35 % of the resolved antonym problems were based on correct mapping information against 76 % of the resolved hyponym problems. Further, among the problems where the expected answer is obtained, only 40 % of antonym problems are based on correct mapping information against 85 % hyponym problems. Therefore, the authors conclude that the information about adjectives in SUMO and its mapping is not suitable for reasoning purposes.

3. Knowledge Framework

For our research purposes, the language resource we use is the Multilingual Central Repository (MCR) (Gonzalez-Agirre et al., 2012), a repository that integrates wordnets from six different languages: English, Spanish, Catalan, Basque, Galician and Portuguese in the same EuroWordNet framework. Additionally, it also integrates other language resources such as Adimen-SUMO (Álvarez et al., 2012), the Top Ontology (Rodríguez et al., 1998) and the Basic Level Concepts (BLC) (Izquierdo et al., 2007). In the MCR adjectives are characterized as in the English WordNet, but they are related to other PoS via the relations *pertainym*, *related*, and *xpos*. For brevity, we will use henceforth *related* to refer to the aforementioned three relations interchangeably.

In this paper, we study a subset of adjective-noun clusters and their corresponding antonyms. As a starting point, we have decided to focus on clusters whose nouns are the hyponyms of the synset *quality_n¹*, which is according to WordNet “an essential and distinguishing attribute of something or someone”. *quality_n¹* is the most frequent hypernym in the adjective-noun clusters and as BLC, it has 1,352 descendants. According to the mapping to WordNet, *quality_n¹* is subsumed by the SUMO class *Attribute*. To sum up, there are 3,802 pairs of antonym adjectives in WordNet and 204 of those pairs appear in the studied adjective-noun clusters. In addition, the two adjective synsets are connected to the same SUMO concept in 934 antonym pairs of WordNet, from which 55 pairs appear in the studied adjective-noun clusters. Thus, we have considered around a 5 % of the adjectives in SUMO.

3.1. Adjective clusters under Quality

We characterized the *quality* adjective-noun clusters in four different types.

The first type (as in Figure 2) is a four-sided cluster where antonym adjectives are related to antonym nouns, which are hyponyms of $quality_n^1$. In this example, the adjective $changeable_a^2$ is related to the synset $changeability_n^1$ and is antonym of the adjective $unchangeable_a^1$. At the same time, $unchangeable_a^1$ is related to $unchangeability_n^1$, which is antonym of $changeability_n^1$. Both nouns have the same hypernym, $quality_n^1$. We represent this cluster in Figure 2.

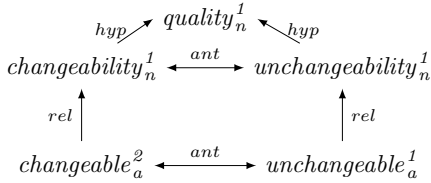


Figure 2: Example of a four-sided cluster

In this case, the three nominal synsets $quality_n^1$, $changeability_n^1$ and $unchangeability_n^1$ are subsumed by the SUMO class *Attribute* while the adjective synsets $changeable_a^2$ and $unchangeable_a^1$ are subsumed by *capability_r*. Obviously, the current knowledge encoded in both WordNet and SUMO do not allow to infer that these qualities (being nouns or adjectives) refer to the capacity or incapacity of things to change. In fact, this cluster should be related somehow to the verbal synset $change_v^1$. Additionally, the SUMO concepts associated to the synsets of the cluster also require a more specific characterization and axiomatization to perform a proper inference about this quality.

These clusters, moreover, can have more than one level due to the hyperonymy. In Figure 3, we show a four-sided cluster with two levels of hyperonymy (second type).

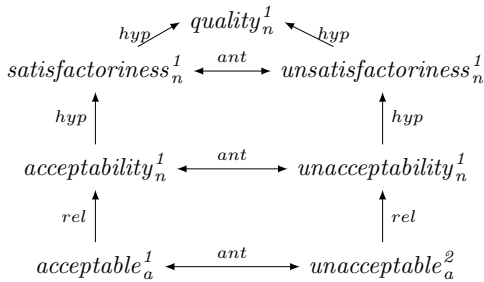


Figure 3: Example of a four-sided cluster with various levels of hyponymy

In this case, the nominal synsets $satisfactoriness_n^1$ and $acceptability_n^1$ and its antonyms are subsumed by the SUMO class *SubjectiveAssessmentAttribute* while the adjective synsets $acceptable_a^1$ is subsumed by the SUMO class *SubjectiveWeakPositiveAttribute* and $unacceptable_a^2$ is subsumed by *SubjectiveStrongNegativeAttribute*. Again, the current knowledge encoded in both WordNet and SUMO

is not sufficient for a proper reasoning about this quality. Moreover, the SUMO classes *SubjectiveWeakPositiveAttribute* and *SubjectiveStrongNegativeAttribute* are not incompatible in SUMO.

The third example of cluster is illustrated in Figure 4. In this case, both adjectives $able_a^1$ and $unable_a^1$ are related to the noun $ability_n^1$, which is an hyponym of $quality_n^1$, forming a three-sided cluster.

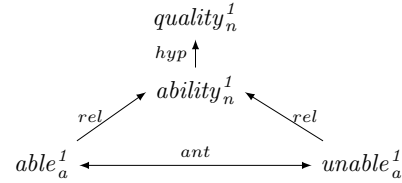


Figure 4: Example of a three-sided cluster

In this case, the nominal synset $ability_n^1$ is subsumed by the SUMO class *Attribute* while the adjective synsets $able_a^1$ and $unable_a^1$ are subsumed by the SUMO relation *capability_r*. Again, the current knowledge encoded in both WordNet and SUMO is not sufficient for a proper reasoning about abilities.

And, finally, the fourth case is presented in Figure 5, a three-sided cluster with an hyponymy chain in one side.

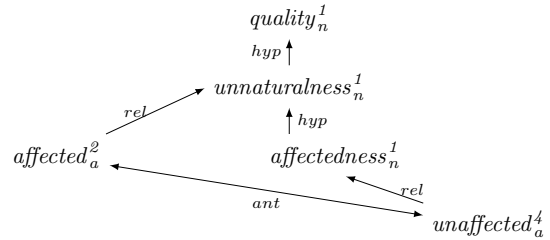


Figure 5: Example of a three-sided cluster with hyponymy

Similarly to previous examples, the nominal synset $unnaturalness_n^1$ and $affectedness_n^1$ are subsumed by the SUMO class *SubjectiveAssessmentAttribute* while the adjective synsets $affected_a^2$ and $unaffected_a^4$ are subsumed by the SUMO class *Pretending_c*. Again, the current knowledge encoded in both WordNet and SUMO is not sufficient for a proper reasoning about this behaviour.

In Table 1 we present the number of clusters per type presented above. In total there are 263 adjective clusters associated to $quality_n^1$ involving 359 adjective synsets and 302 nominal synsets.

Cluster type	Cluster Number
Four-sided clusters	51
Four-sided clusters (various levels)	102
Three-sided clusters	98
Three-sided clusters (various levels)	12
Total	263

Table 1: Number of Cluster for Type

We have also detected that some clusters are not fully-formed. That is, some antonym relations between the adjectives or between the nouns are missing. These incomplete clusters will be studied in a near future.

4. Improving the knowledge framework

Being one of our main motivation to reason with SUMO properties, we need to properly augment the ontology with new knowledge related to qualities from WordNet and, on the other hand, we need to correctly map the quality clusters to the ontology.

4.1. Improving WordNet relations for qualities

Inspired by the WordNet morphosemantic links, the idea is to create new semantic relations between synsets in a cluster to the corresponding nouns and verbs they are related, if possible. The morphosemantic links took into account English morphology to create the relation. In this work, on the one hand, we have taken more derivative relations into account, and, on the other hand, we also have considered the morphology of the latinate borrowings. For example, we have linked the adjective *impalpable*_a¹ and the noun to which is already related, *impalpability*_n¹, to the verb *touch*_v². This work has been done manually taking into account the following guidelines. For brevity, we only use one member of the cluster in the examples.

- Link the nouns and the adjectives in the cluster to the synset with the most general meaning e.g. *advisable*_a¹ “worthy of being recommended or suggested; prudent or wise” to the verb *advise*_v¹ “give advice to”.
- In case of the *ambiguous* clusters, link to all the possible synsets. For example, the adjective *comprehensible*_a¹ “capable of being comprehended or understood” to the verbs *understand*_v¹ “know and comprehend the nature or meaning of” and *comprehend*_v¹ “get the meaning of something”.
- In case of clusters with various levels and repeated hypernyms, keep the link to the same synset if possible. For instance, in the clusters with *changeability*_n¹ as hypernym that includes in other levels respectively the adjectives *variable*_a¹, *mutable*_a¹, *alterable*_a¹ among others are linked to the same verb: *alter*_v¹ “cause to change; make different; cause a transformation”.
- Do not link if there is no right sense e.g. *auspicious*_a¹ cf. Spanish *auspicar* or Italian *auspicare* evolved from Latin *auspicium* and *auspicare*.

This way, we have created 233 new *quality_of* relations, 139 for events and 94 for nouns. Henceforth, we will denominate *the top synset of the cluster* the synset (the noun or the verb) they are linked to, i.e. the concept/event whose qualities they express.

However, 69 clusters could not be related to any noun or verb and these have been marked as pure. An example of this is the cluster that contains the adjectives *good*_a¹ <-> *bad*_a¹ and the nouns *goodness*_n¹ <-> *badness*_n¹.

4.2. Grouping clusters under quality

As a result of the new relations, we organized the synset clusters under *quality*_n¹ as follows:

- **Qualities of Events:** These are clusters related to qualities of verbs. For instance, the cluster *changeable*_a² [*changeability*_n¹] <-> *unchangeable*_a¹ [*unchangeability*_n¹] denotes qualities related to the verb *change*_v¹ (see Figure 2). There are 107 clusters related to events.
- **Qualities of Nouns:** These are clusters related to concrete and abstract nouns. For example, the clusters including the hypernyms *faithfulness*_n¹ and *humanness*_n¹ have been related respectively to *faith*_n⁴ “loyalty or allegiance to a cause or a person” and *person*_n¹ “a human being”. There are 86 clusters related to nouns.
- **Pure Qualities:** In this case, the members of these clusters cannot be linked to verbs or nouns and we have marked them as *pure* e.g. *bad*_a¹ and *badness*_n¹. There are 70 clusters classified as pure.

These groupings are the basis for the ontologisation model.

4.3. New top ontology for qualities

As we are working with qualities, we select the class of attributes *Attribute*, whose semantics—according to SUMO documentation—is “Qualities which we cannot or choose not to reify into subclasses of”, as super-concept of all the new defined concepts. Since the hypernym of all the considered clusters is *quality*_n¹, the first new concept we propose in our model is *QualityAttribute*, which is defined as a subclass of the SUMO class *Attribute*. *QualityAttribute* is the top class of the model constructed for the considered clusters.

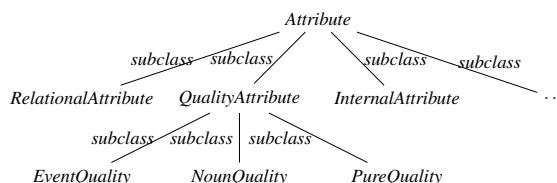


Figure 6: New ontology model for qualities

According to the created subtypes of qualities, we define three new direct subclasses of *QualityAttribute*: *EventQuality*, for qualities of events; *NounQuality*, for qualities of nouns; and *PureQuality*, for pure qualities (see Figure 6).

4.4. Integrating the clusters and the new ontology

Further, we create a new class of attributes for each top synset of the cluster, which is defined as direct subclass of *EventQuality*, *NounQuality* or *PureQuality* according to the subtype of the top synset. Hence, we introduce 61 new subclasses of *EventQuality*, 45 new subclasses of *NounQuality* and 32 new subclasses of *PureQuality*. The labels of these new classes are formed by capitalizing the first letter of the wordform of the top synset of the cluster and appending

the string *Quality*. This way, *ChangeQuality* has been created on the basis of the synset $change_v^1$ by converting it to *Change* and concatenating *Quality*. From now on, we will refer to the class created for the top synset of a cluster as the *cluster class*.

On the basis of the proposed new ontology of qualities, we obtain a new mapping for the nouns and adjectives in the clusters by using the *equivalence* mapping relation and its complementary. For this purpose, we automatically connect the antonym pairs of noun and adjectives synsets of a given cluster to its cluster class, but with opposite semantics: that is, given a pair of antonym synsets in a cluster where A is the cluster class, one of the antonym synsets is stated to be related with A by *equivalence*, while the other one is stated to be related with A by the complementary of *equivalence*. For simplicity, from now on we say that the polarity of a synset is *positive* if it is related with the corresponding cluster class by *equivalence*, and it is *negative* otherwise (related with the complementary of *equivalence*). When we refer to polarity in this paper we do not take into account the polarity of the concept, but the polarity of the word: if the attribute is present or not. That is, *fear* can be understood as a negative concept, and *fearless* as a positive, but in this paper, *fear* is positive in the sense that the attribute *fear* is present and *fearless* is negative because it implies that there is no *fear*.

In order to automatically decide the polarity of the antonym synsets in a cluster, we analyze the senses of the involved synsets in the following way: given two antonym synsets with senses s_1 and s_2 respectively such that s_2 is substring of s_1 ,

- If either “*a-*”, “*de-*”, “*dis-*”, “*il-*”, “*im-*”, “*in-*”, “*ir-*”, “*mis-*”, “*non-*” or “*un-*” is prefix of s_1 , then the polarity of s_1 is *negative* and the polarity of s_2 is *positive*.
- Else if “*-less*” is suffix of s_1 , then the polarity of s_1 is *negative* and the polarity of s_2 is *positive*.
- Otherwise, the polarity of s_1 and s_2 is unknown.

For example, let us consider the cluster in Figure 2. Since the “*changeable*”/“*changeability*” are substring of “*unchangeable*”/“*unchangeability*”, which has “*un-*” as prefix, then the polarity of “*changeable*”/“*changeability*” is *positive* while the polarity of “*unchangeable*”/“*unchangeability*” is *negative*. Consequently, “*changeable*”/“*changeability*” are connected to *ChangeQuality* by *equivalence* while “*unchangeable*”/“*unchangeability*” are connected to *ChangeQuality* by the complementary of *equivalence* in the new proposed mapping.

However, the above mentioned heuristics cannot be applied in some clusters because the polarity of the antonym synsets is unknown. In this case, we create two new classes of attributes, which are defined as contrary each other and subclass of the cluster class. This enables to state that each synset from antonym pairs are related to incompatible classes of attributes and, this way, the process of mapping the antonym nouns and adjectives of the considered clusters is fully automatic. For example, the antonym nouns $difficult_n^1$ and $simpleness_n^3$ and antonym

adjectives $difficult_a^1$ and $easy_a^1$ form a four-sided cluster (first type) with *DifficultyQuality* —which is subclass of *NounQuality*— as cluster class, but the polarity of the antonym synsets cannot be automatically decided by our proposed method. To overcome this problem, we create two new contrary classes of attributes, *DifficultyQuality* and *SimplenessQuality*, which are defined as subclass of *DifficultyQuality*. Thus, in the resulting mapping, $difficult_n^1$ and $difficult_a^1$ are related with *DifficultyQuality* by *equivalence* and $difficult_n^1$ and $easy_a^1$ are related with *SimplenessQuality* by also *equivalence*. This way, we create 29 pairs of new contrary classes (that is, 58 new classes) distributed as follows: 8 new subclasses of *EventQuality*, 24 new subclasses of *NounQuality* and 26 new subclasses of *PureQuality*.

5. Validation

In this section, we summarize and validate the result of our proposal for the new ontology for qualities, the new WordNet relations and the new SUMO mapping to WordNet adjectives.

In total, we have augmented SUMO by introducing 200 new classes of attributes, which have been defined as subclass of *Attribute*. For their axiomatization, we have stated that 41 pairs of attribute classes are contrary of each other. Using the new axiomatization, we have successfully connected 722 synsets: 61 verbs, 302 nouns and 359 adjectives. Further, the mapping of the adjectives can be propagated to another 1,384 satellite adjectives by using the *similarity* relation.

We have also checked the suitability of the resulting mapping. More specifically, we have verified that all the antonym pairs have an incompatible mapping between each other. Consequently, the new proposed ontology and mapping can be applied in commonsense reasoning tasks involving WordNet adjectives.

6. Conclusion and Future Work

In this paper we have presented the first steps towards modeling the attributes expressing qualities in SUMO based on the knowledge encoded in WordNet. To that end, in this experimental sample, we have focused on studied the clusters of adjectives and nouns related to the synset $quality_n^1$. When necessary, we have related the clusters to the corresponding nominal and verbal qualities. Based on these relations, we have created new classes in the ontology and we have mapped the synsets to them.

For the future, we plan to explore how to spread this approach as automatically as possible. First we want to study the non fully formed clusters (those that have a missing relations), and other adjective types such as those denoting properties. We also plan to explore other options or resources to associate the polarity to synsets (Agerri and García-Serrano, 2010). Moreover, we foresee to test the model and the added information in a commonsense reasoning system relating properties.

7. Acknowledgements

This work has been partially funded by the project Deep-Reading (RTI2018-096846-B-C21) supported by the Min-

istry of Science, Innovation and Universities of the Spanish Government, and GRAMM (TIN2017-86727-C2-2-R) supported by the Ministry of Economy, Industry and Competitiveness of the Spanish Government, the Basque Project LoRea (GIU18/182), Ixa Group-consolidated group type A by the Basque Government (IT1343-19) and BigKnowledge – *Ayudas Fundación BBVA a Equipos de Investigación Científica 2018*.

8. Bibliographical References

- Agerri, R. and García-Serrano, A. (2010). Q-WordNet: Extracting Polarity from WordNet Senses. In *LREC*.
- Álvez, J., Lucio, P., and Rigau, G. (2012). AdimenSUMO: Reengineering an ontology for first-order reasoning. *Int. J. Semantic Web Inf. Syst.*, 8(4):80–116.
- Álvez, J., Gonzalez-Dios, I., and Rigau, G. (2019a). Commonsense reasoning using WordNet and SUMO: a detailed analysis. In *Proc. of the 10th Global WordNet Conference (GWC 2019)*, pages 197–205.
- Álvez, J., Lucio, P., and Rigau, G. (2019b). A framework for the evaluation of SUMO-based ontologies using WordNet. *IEEE Access*, 7:36075–36093.
- Fellbaum, C., Gross, D., and Miller, K. (1993). Adjectives in wordnet. In George A. Miller, editor, *Five Papers on WordNet*, pages 26–39. Technical Report, Cognitive Science Laboratory, Princeton University.
- Fellbaum, C., Osherson, A., and Clark, P. E. (2007). Putting semantics into wordnet’s ‘morphosemantic’ links. In *Language and technology conference*, pages 350–358. Springer.
- C. Fellbaum, editor. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.
- Gangemi, A., Nuzzolese, A. G., Presutti, V., and Reforgiato Recupero, D. (2016). Adjective Semantics in Open Knowledge Extraction. In *Formal Ontology in Information Systems: Proceedings of the 9th International Conference (FOIS 2016)*, volume 283, page 167. IOS Press.
- Gonzalez-Agirre, A., Laparra, E., and Rigau, G. (2012). Multilingual Central Repository version 3.0. In N. Calzolari, et al., editors, *Proc. of the 8th Int. Conf. on Language Resources and Evaluation (LREC 2012)*, pages 2525–2529. European Language Resources Association (ELRA).
- Hamp, B. and Feldweg, H. (1997). GermaNet-a Lexical-Semantic Net for German. In *Automatic information extraction and building of lexical semantic resources for NLP applications*.
- Henrich, V. and Hinrichs, E. (2010). GernEdiT-the GermaNet editing tool. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*, pages 2228–2235.
- Izquierdo, R., Suárez, A., and Rigau, G. (2007). Exploring the Automatic Selection of Basic Level Concepts. In *Proceedings of the International Conference on Recent Advances on Natural Language Processing (RANLP’07)*, volume 7.
- Niles, I. and Pease, A. (2001). Towards a standard upper ontology. In Guarino N. et al., editor, *Proc. of the 2nd Int. Conf. on Formal Ontology in Information Systems (FOIS 2001)*, pages 2–9. ACM.
- Niles, I. and Pease, A. (2003). Linking lexicons and ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In H. R. Arabnia, editor, *Proc. of the IEEE Int. Conf. on Inf. and Knowledge Engin. (IKE 2003)*, volume 2, pages 412–416. CSREA Press.
- Peters, I. and Peters, W. (2000). The Treatment of Adjectives in SIMPLE: Theoretical Observations. In *LREC*.
- Rodríguez, H., Climent, S., Vossen, P., Bloksma, L., Peters, W., Alonge, A., Bertagna, F., and Roventini, A. (1998). The top-down strategy for building EuroWordNet: Vocabulary coverage, base concepts and top ontology. In *EuroWordNet: A multilingual database with lexical semantic networks*, pages 45–80. Springer.
- Tsvetkov, Y., Schneider, N., Hovy, D., Bhatia, A., Faruqui, M., and Dyer, C. (2014). Augmenting english adjective senses with supersenses. In *Proc. LREC’14*.

Incorporating Localised Context in Wordnet for Indic Languages

S. H. Mohapatra¹, S. Agnihotri¹, A. Garg, P. P. Shah, S. Chakraverty

Netaji Subhas University of Technology

Dwarka, New Delhi

{soumyam, shikhara, aparg, praveens}.co.16@nsit.net.in, shampa@nsit.ac.in

Abstract

Due to rapid urbanization and a homogenized medium of instruction imposed in educational institutions, we have lost much of the golden literary offerings of the diverse languages and dialects that India once possessed. There is an urgent need to mitigate the paucity of online linguistic resources for several Hindi dialects. Given the corpus of a dialect, our system integrates the vocabulary of the dialect to the synsets of IndoWordnet along with their corresponding meta-data. Furthermore, we propose a systematic method for generating exemplary sentences for each newly integrated dialect word. The vocabulary thus integrated follows the schema of the wordnet and generates exemplary sentences to illustrate the meaning and usage of the word. We illustrate our methodology with the integration of words in the Awadhi dialect to the Hindi IndoWordnet to achieve an enrichment of 11.68 % to the existing Hindi synsets. The BLEU metric for evaluating the quality of sentences yielded a 75th percentile score of 0.6351.

Keywords: IndoWordnet, geographical wordnets, lexicons, clustering

1. Introduction

The Hindi Belt or Hindi heartland, is a linguistic region consisting of parts of India where Hindi and its various dialects are spoken widely (Sukhwai, 1985). Hindi as a language has evolved over the years due to migration and invasion of various socio-ethnic groups like Turks, Britishers etc. This has given it a dynamic shape with the Devanagari script remaining nearly the same but the speech changing with location, thereby leading to a plethora of dialects spread across the region. Of late, due to westernisation and globalization, over 220 Indian languages have been lost in the last 50 years, with a further 197 languages marked as endangered according to People’s Linguistic Survey of India, ‘18 (V. Gandhi, 2018). There is thus an exigent need to preserve not only the Hindi language but its various dialects that give India its unique identity embodying unity in diversity. Through this project we take a step towards protecting this linguistic heritage.

The Indo-wordnet is a linked structure of wordnets of major Indian languages from the Indo-Aryan, Dravidian and Sino-Tibetan families. It was created by following the expansion approach from Hindi wordnet which was made available free for research in 2006 (Bhattacharya, 2006). However, each Indic language has a number of dialects for which the IndoWordnet has no related information. In this paper, we enhance this digital footprint by systematically incorporating vocabulary from Hindi dialects using language processing tools, algorithms and methods.

Our research contributes by first collating the resources of a dialect that are available in Devanagari script from multiple textual sources as well as from audio clips of real conversations. This consolidated corpus is subsequently used to extract the vocabulary and exemplify

its appropriate usage to enrich the indowordnet. Ultimately, the wordnet is envisioned to be a complex whole containing not just word usages but the way a particular word is pronounced and used in different geographical regions. We demonstrate our methodology by using the Awadhi dialect to enrich the Hindi IndoWordnet.

2. Prior Work

Of late, several research groups have contributed towards enrichment of wordnet for different languages.

Researchers from Jadavpur University (Ritesh, 2018) have developed an automatic language identification system for 5 closely-related Indo-Aryan languages of India namely, Awadhi, Bhojpuri, Braj, Hindi and Magadhi. They have compiled corpora with comparable format but varying lengths for these languages by tapping upon various resources.

Mikhail et al. (Mikhail, 2017) present an unsupervised method for automatic construction of WordNets based on distributional representations of sentences and word-senses using readily available machine translation tools. Their approach requires very few linguistic resources and can thus be extended to multiple target languages.

Nasrin Taghizadeh et al. (Nazrin, 2016) propose a method to develop a wordnet by only using a bi-lingual dictionary and a mono-lingual corpus. The proposed method has been executed with Persian language. The induced wordnet has a precision of 90% and a recall of 35%.

Taking inspiration from the above approaches, we formulate our own approach and propose an algorithm to build wordnets for low resource dialects of Hindi. Our work primarily focuses on dialects, an area which has thus far been ignored.

¹ These authors have contributed equally

3. Method

We present an algorithm that takes in the corpus of a given dialect and its Hindi bilingual dictionary. The system uses this dialect's vocabulary to enhance the synsets of Hindi IndoWordnet. Our twin goals are to ensure that the enriched vocabulary follows the schema of the wordnet and that we are able to generate exemplary sentences to illustrate the meaning and usage for each dialect word.

3.1 Data Collection

We used the Awadhi-Hindi bilingual dictionary, also called *Awadhi Shabdkosh*², an ebook, for extracting Awadhi words along with their relevant Hindi meanings. We used this collection solely for the purpose of IndoWordnet synset integration.

The main source of our corpus compilation came from the comparable corpora from Jadavpur University (Ritesh, 2018) which consists of a training set of 70350 lines, a validation set of 10300 lines, test data of 9600 lines and 9600 lines of gold data for test sentences. The gold data contains the labels for the test data. The comparable dataset consists of tagged sentences belonging to Awadhi, Bhojpuri, Braj, Hindi and Magadhi categories.

A total of 12297 Awadhi sentences were extracted from the Jadavpur corpus on the basis of these tags. We also found other sources of Awadhi literature in electronic form including an ebook containing Bible Stories in Awadhi and audio samples of conversations on social topics³. These additional resources yielded 3500 Awadhi sentences. The consolidated dataset is used for creating an Awadhi lexicon of preferential pairs of Awadhi words and for sentence generation.

3.2 Automatic Mapping of Existing Resources

IndoWordnet provides the NLP resources of various Indic languages. However, it does not store any linguistic information about the various dialects in which a word may be spoken in different regions. In this section, we explain the processes involved in mapping the Awadhi words in the bilingual dictionary to the relevant Hindi synsets.

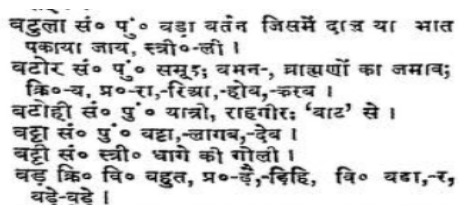


Figure 1: Image from the *Awadhi Shabdkosh*: The Awadhi-Hindi bilingual dictionary

3.2.1 Preparation of Inverse Dialect Mappings

An excerpt from the *Awadhi Shabdkosh* is shown in Figure 1. An awadhi word (say, बटोर) is followed by its POSTag (सं for noun), its gender (पुं for male) and its meaning in hindi. The symbol '।', known as the purnaviram, marks the end of a sentence.

Using an Optical Character Reader⁴ (OCR) and regex, we bring the PDF data into usable textual format.

We pick each Indic word from the inverted mappings and search it in the IndoWordnet. If it exists then we integrate the dialect word into the first synset of the Indic language and add a metadata tag to the same indicating the dialect to which it belongs. The regex used to generate inverted mappings is shown in Figure 2.

```
Regex = r“( [०पु०]([\u0900-\u0965\u0967-\u097F]+
[.;:|;|([\u0900-\u0965\u0967-\u097F]+०) ”g
```

Figure 2: Regex used to extract mappings from plain text

The regex helps us to extract pairs of Awadhi-Hindi words with each Awadhi word having a single word Hindi meaning so as to preclude superfluous words. Each regex match has 3 groups. The green group picks up the POSTag delimiter, the blue group picks up the single Hindi word we are concerned with and the red group picks up the delimiter for the Hindi word. The Hindi word delimiter group helps filter out single word meanings.

Figure 3, shows different matches detected using the regex 101 tool⁵. To illustrate, in the first match, the Awadhi word is लड़कपिल्ली, the POSTag delimiter is ०, the Hindi word is लड़खड़ाब and Hindi word delimiter is किर०. Hence, the Awadhi-Hindi pair is लड़कपिल्ली - लड़खड़ाब.

3.2.2 Using Metadata Tags

With the integration of dialect related lingual information for each Indic language in the IndoWordnet, it becomes mandatory to add metadata tags to each word to indicate its membership in different dialects. Adding metadata tags enriched the semantic information of each word in the IndoWordnet.

For every new word that is being integrated to the IndoWordnet, the metadata tag denotes the dialect from which the request for integration into the wordnet has been instantiated. Figure 4a shows an Awadhi word with its Hindi meaning. Figure 4b shows the results after integration of Awadhi word to the Hindi synset along with an AWD tag, showing membership to the Awadhi dialect.

3.3 Knowledge Representation

This section documents the different data representations we developed out of the existing data.

² archive.org/details/in.ernet.dli.2015.481490/mode/1up

³ <https://bit.ly/2ySzXWY>

⁴ <https://pypi.org/project/google-cloud-vision/>

⁵ <https://regex101.com/>

लड़कपिल्ली, "वि० पु० चिबिल्ला लड़का ; वै० लड़खड़ाब किर० सं० हिलकर गिरने लगना ;
लड़हरा, सं० पु० घरी का लंबा पेड़ ।
लड़ाइब, दे० लड़य । ।
लड़ाई, "सं० स्त्री० सुद्ध , झगड़ा करब , होय ।"
लड़ाका, वि० झगड़ालू ।

Figure 3: Regex matches as seen on Regex 101 web tool

पार सं० पु० किनारा; पाइब, जीतना, करब, होब;
-लागब, हो सकना; लगाइब ।

Figure 4a: Awadhi word पार and its Hindi meaning किनारा in Awadhi-Hindi bilingual dictionary

Hindi word/URL	Old Synset	New Synset
किनारा	किनारा, किनार, कोर, सिरा, छोर, उपांत, अवारी, आर, पालि, झालर	किनारा, किनार, कोर, सिरा, छोर, उपांत, अवारी, आर, पालि, झालर, पार(AWD)

Figure 4b: Hindi word (किनारा) along with its old and new synset

3.3.1 Stop Words and POS Tags

Since there is no pre-existing list of stop words for Awadhi, we make our own. We first create a frequency map of all the words in the corpus and sort them in descending order of frequency of occurrence. Words which have a high frequency, and in addition belong to the class of determiners, prepositions or conjunctions such as यह (this), उसका (their) and और (and) are added to the list of stop words.

We use the Hidden Markov Model (HMM) to POS tag our precompiled Awadhi dataset. The HMM requires a set of observations and a set of states. Words in a sentence are defined as the observations and the POS tags are the hidden states. The HMM uses a transition probabilities matrix and a conditional probabilities matrix. For a given pair of POSTags, say (ADJ and NP) the transition probability TP is defined by the conditional probability:

$$TP = P(\text{ADJ} | \text{NP}) \quad \dots \text{Eq 1}$$

For a given word a and POSTag, the emission probability EP is defined by the conditional probability:

$$E.P. = P(a | \text{NP}) \quad \dots \text{Eq 2}$$

In order to build the two matrices, we use pre-tagged hindi dataset available from the Universal Dependencies, UD_Hindi-HDTB dataset (Riyaz, Martha, 2009). This dataset consists of close to 2000 POSTagged sentences in the Hindi language. Once we train this model, for a new sentence it uses the pre-built matrices to predict the POSTags.

Figure 5 shows the result of POS tagging a Awadhi sentence using the HMM based POS tagger. The English translation of the sentence is - "Being a minister it becomes his duty to listen to both the parties".

3.3.2 Lexicon

We create a lexicon of concept words (nouns) and preferential word pairs with the help of the POSTagged Awadhi dataset. This lexicon serves as a rich source of conceptually cohesive words to build sentences with improved factual correctness.

Input : परधान होय के नाते उनका दुनहू तरफ कि सुनैक जरूरी रहै .
Output : परधान/PROPN होय/PROPN के/ADP नाते/ADP
उनका/PRON दुनहू/NOUN तरफ/ADP कि/SCONJ
सुनैक/NOUN जरूरी/ADJ रहै/VERB ./PUNCT

Figure 5: Result of POSTagged Awadhi sentence

We pick up nouns from the POS tagged dataset. We plot a graph of NP (noun phrases) identified, based on their word embeddings. Each NP serves as a node and the edge weight is the inverse of the cosine distance between the word embeddings. We generate clusters of the plotted nodes. A dense cluster signifies a set of nouns which are used together frequently and hence represent a conceptually cohesive set. Thus, we pick up the cluster having the highest number of nouns. These NPs serve as the final set of nouns that are included in our lexicon.

We now build the rest of the lexicon. From the dataset, we first allot each ADJ a proximity score based on the number and the closeness of the selected NP around it. We pick a set of top 'n' unique adjectives, based on their scores. These will now serve as the final set of preferential noun-adjective pairs in our lexicon. We perform the same sequence of steps to pick up preferential noun-verbs, noun-pronoun and verb-adverb pairs in the lexicon.

For example, let the noun word be प्रकृति (nature) for which the corresponding adjectives are अनगिन्त (limitless), सहज (spontaneous), कृतघ्न (not showing gratitude), ममतामयी (mother's kindness) and स्थायी (fixed, not changing).

3.3.3 Digital Dictionary

The inverse dialect mappings created to enrich the present Indowordnet (refer subsection 3.2.1) also serves as a resourceful bilingual dictionary in digital form for a given dialect word. Using our sentence generation model explained in the next subsection, we further enrich this dictionary with example sentences for each word-meaning pair.

3.4 Sentence Generation using Recurrent Neural Networks

We designed a Recurrent Neural Network (RNN) that helps us in generating meaningful sentences using a dialect word as seed. Since Awadhi is a low resourced language, RNN is seen as a good method for sentence generation (Gandhe, 2014). The sentences serve as exemplary sentences for the newly added Awadhi word to the IndoWordnet. The motivation for using RNN rather than pick up an example sentence from the corpus itself is to be able to generate new sentences that highlight the local cultural aspects of the dialect. This aligns with our objective of preserving heritage and we will address this issue as the next step in our project's roadmap.

Alternatively we also used the N-gram model to generate sentences. This model takes in a set of words and generates a score of each possible permutation based on Markov probability rules (Yadav, 2014). The limitation of this model lies in its prerequisite to provide the entire list of words that the sentence would comprise of. Furthermore, it cannot construct sentences from a single seed word as is possible with RNN. It is noteworthy that even though RNN has been used for sentence construction in Bangla (Islam, 2019) and English (Sutskever, 2011), it is for the first time that it has been used for sentence generation in a Hindi dialect - Awadhi.

We trained the RNN on the set of Awadhi sentences compiled in our corpus from multiple sources as mentioned in section 3.1. The RNN aims at understanding the syntactic constructs of words in a sentence so that it can use this knowledge in predicting words that are most probable in a given context.

To ensure that the sentences being generated are semantically correct, we make use of the preferential word pair lexicon we developed in subsection 3.3.2. During each step of next-word prediction in RNN, the model returns an array of probabilities for the next word. Using the lexicon relations we selectively nullify the probability scores of unrelated words. For example, for the root word - *पिता*, the top 5 words with highest probabilities in the probabilities array returned by the RNN are [*जी*, *जान*, *कठोर*, *अपनी*, *पिरय*]. The noun-adjective lexicon pair for *पिता* is [*पुलकित*, *परम*, *कठोर*, *साध्वी*, *पिरय*]. Hence, after nullifying the probabilities of the words not present in the lexicon, the top 5 words now are [*कठोर*, *पिरय*, *गुरु*, *तुल्य*, *परम*].

Our training set consists of “s:t” pairs that correspond to a list of 5 words in sequence and the next-word in sequence respectively. Figure 6 below shows (s) as an array of 5 words in sequence and (t) as the next word in this context.

```
['\n', 'हमरे', 'लिए', 'इतनी', 'जगह'], next_word: काफी
['बहिका', 'रंग', 'ओ', 'मुखादि', 'फूलमती'], next_word: कि
['बने', 'सेनी', 'अत्ती', 'तपनि', 'मैहा'], next_word: खुब
['तेरह', 'मा', 'यहि', 'पोर्टल', 'पर'], next_word: जेतना
```

Figure 6: Training pairs generated from Awadhi data-set available

To illustrate the process, consider the first training pair. For 5 words - i. \n (newline) ii. हमरे (us) iii. लिए (for) iv. इतनी (this) v. जगह (place) - in sequence the next word is काफी (enough). This “s:t” pair has been extracted from the sequence - \n हमरे लिए इतनी जगह काफी है । (This place is enough for all of us.)

In the fourth training pair, for 5 words - i. तेरह (thirteen) ii. मा (in) iii. यहि (this) iv. पोर्टल (portal) v. पर (on) - in sequence the next word is जेतना (specific). This “s:t” pair has been extracted from the sequence - कुछ ब्यस्तता के चलते दुइ हजार तेरह मा यहि पोर्टल पर जेतना काम हुवे क रहा, नाय होइ पावा । (Specific work on this portal couldn't be completed due to some busy schedules in two thousand thirteen.)

Of the available Awadhi sentences for training purposes we use 20% of the dataset for the purposes of validation and the rest for training. Awadhi as a dialect is low resourced and most of the resources available online overlap in their content. During training, we allowed overfitting of the model over the consolidated training set of Awadhi sentences. We observe that overfitting of the data actually helps us to retain the semantic and syntactic relationships between words in the way they occur in the actual text. However, this also leads to a decrease in the overall generalizability of the process of sentence generation.

3.5 Evaluation of Sentence Quality - BLEU

BLEU (BiLingual Evaluation Understudy) is an algorithm for evaluating the quality of machine-translated text from one natural language to another (Kishore, 2002). The BLEU score has been used for measuring machine translated English to Hindi sentences (Malik, 2016). It can also be used for evaluation of sentence similarity.⁶

For each sentence that is generated by the RNN model for a given root word, we create a reference set using the consolidated Awadhi dataset. The reference set selectively contains only those sentences which contain the root word. Choosing sentences which contain the root word ensures that only relevant sentences are compared against and this decreases the chances of getting low scores.

The BLEU model now evaluates the cumulative n-gram scores of the candidate sentence with respect to the reference set, at all orders from 1 to n. NLTK's BLEU model by default calculates the 4-gram cumulative score, with n being set to 4 and the default weights being (0.25, 0.25, 0.25, 0.25). The algorithm finally returns the weighted geometric mean score for all n-gram scores. We make use of this model to evaluate the quality of our generated sentences⁷.

An example is shown in figure 8, where the awadhi root word is जंगल (forest) and the RNN generated sentence is जंगल (forest) मा (in) एक (one) जगहों (place or area) आम (mango) पाक (ripe) रहा. (There is one place in the forest where mangoes are ripening). The reference set

⁶ bit.ly/2V640ms

⁷ bit.ly/3aadZLC

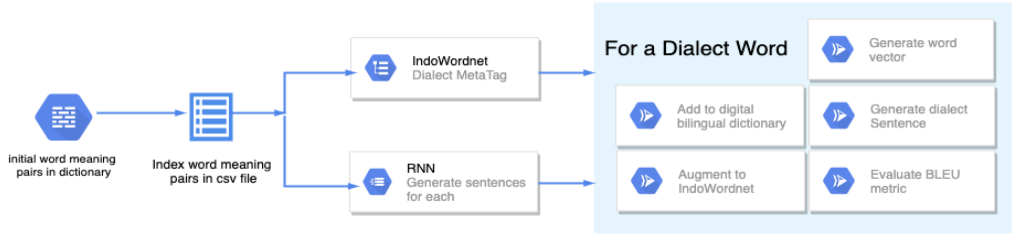


Figure 7: Complete Workflow

Figure 7 shows the complete workflow, starting from word meanings in *Awadhi Shabdkosh*, inverted word-meaning pairs, integration to IndoWordnet to sentence generation using RNN and sentence evaluation using BLEU

consists of 9 sentences which contain the awadhi word - *जंगल*.

The BLEU score is calculated for the above sentence using the reference set giving us a score of 0.4012.

```
example_reference = [
    "चउगइल जंगल की राह लिहिस",
    "जंगल मा बइठा रहई",
    "सिकार उठा जंगल मा भाग",
    "सब केउ जंगल मा जुटि के गयेनि",
    "साधू जायि के उही जंगल मा फेरि राम राम करइ लागेनि",
    "जब लगे गयेनि तउ देखेनि कि हारमती जंगल के बन्दी खाने मा परी बिलपत बा",
    "बहि कयि रोउब सुनि के जंगल कयि चिरई जुटि गई",
    "सोना बहिनी रोवत की बीच जंगल मा बयिठी",
    "सोना धरे से जायि के जंगले मा रोवयि लागी" ]

generated_sentence = "जंगल मा एक जगहाँ आम पाक रहा "

sentence_bleu(example_reference, generated_sentence)

0.40126711450090535
```

Figure 8: BLEU score for the generated sentence

4. Experimental Results

4.1 IndoWordnet Enrichment

The Awadhi-Hindi bilingual dictionary has 37 alphabets ranging from अ to य. Table 3 shows the total number of Awadhi words under column 3 (TW) that exist under each alphabet - shown under column 1 (CE) in English and column 2 (CH) in Hindi. The total number of Awadhi words having one word Hindi meanings are shown under column 4 (TWSM). The inversion process led to an average of 48.48% loss in Awadhi words collected.

The number of Hindi synsets enriched with their Awadhi equivalents and their exemplary sentences due to the next step of integration to the IndoWordnet is shown under column 5 (SE). This step incurs a further miss rate of 30.91% on an average. This loss was seen to occur due to the following two factors - 1) OCR does not identify the Hindi word in the bilingual dictionary correctly. For example, in figure 9 the OCR interprets खट्टापन (sourness) as खापन (no such word exists). The target Hindi word doesn't exist in the IndoWordnet. For example, the Hindi word मइजिल doesn't exist in IndoWordnet. Figure 10 plots the alphabetical inverse mapping losses and IndoWordnet integration losses.

The IndoWordnet consists of 26,000 synsets for the Hindi language⁸ and the number of synsets enriched due to the Awadhi corpus is 3036 (11.68%). This is significant, keeping in mind the scarcely available Awadhi datasets. We believe this number will increase when we proceed with other dialects of Hindi such as Braj, Rajasthani, Marwari etc.

खटासि संस्त्री० खट्टापन, थोडी खटाई, वैस ।	
1	word meaning ।
48	खटासि संस्त्री० खापन , थोडी खटाई , वैस ।

Figure 9: (a) Awadhi word खटासि and its meaning in *Awadhi Shabdkosh*. (b) OCR interprets खट्टापन (sourness) as खापन (no such word exists).

4.2 BLEU Scores

The threshold score of 0.6351 was decided on the basis of statistics observed over 158 sentences generated, two for each of 79 Awadhi words chosen randomly from our corpus. The statistics are shown in Table 1. We decided to include sentences with BLEU scores above that corresponding to the 75th percentile score.

Max	Min (Non-zero)	Mean
0.9036	0.1119	0.4679
Median	75th Percentile	90th Percentile
0.4324	0.6351	0.7174

Table 1: BLEU Scores

RNN has not been used yet for generating sentences in Awadhi or any other dialects of Hindi. We show exemplary sentences in Table 2 with scores above threshold for the Hindi words माई (mother) and बच्चा (child) along with their English translation.

The english translations were performed manually by us using the awadhi-hindi and hindi-english bilingual dictionaries. It was seen that for a threshold of 0.6351 sentences were syntactically and semantically correct.

⁸ <https://bit.ly/2XB07HW>

sentences for माई	BLEU
माई (mother) तोहरे (your) संकोच (hesitation) के मारे कुछ (nothing) बोलि (spoke) ना <i>(your mother spoke nothing out of hesitation)</i>	0.7788 ✓
माई संस्कार कहां मिली कि हमरे माई केहू खुशी नाहीं देखे नाहीं तौ भर जाय कै धारि अपने साथ चली जा	0.2069 ✗
sentences for बच्चा	BLEU
बच्चा (children) लोग जौ खाना (food) लाये (brought) हैं ऊ खाय (eating) लागै <i>(The children started eating the food they had brought with them)</i>	0.8948 ✓
बच्चा (child), मंत्र (mantra) एकइ (once) बार काम (works) करा थइ <i>(Listen child, this mantra works only once)</i>	0.6514 ✓

Table 2 : RNN generated sentences

We made a visualisation tool for generating new sentences and several other tasks mentioned in section 4. The link is added in the references.⁹

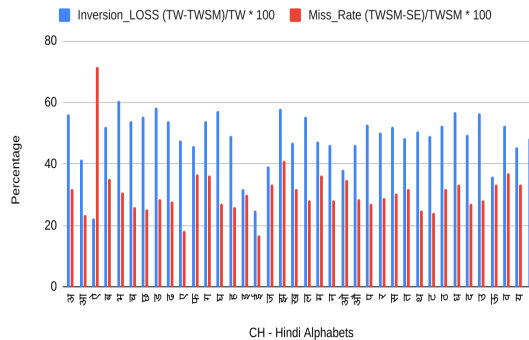


Figure 10: Histogram showing inversion loss and miss percentage.

5. Conclusion and Future Scope

We formulated and presented a methodical and scalable approach to enrich the IndoWordNet. This not only enhances the Indo Wordnet's viability as a social project but also protects local dialects from fading. Similar to the Awadhi dialect, we also plan to expand our approach to the Braj and Marwari dialect. We have bilingual dictionaries for both the dialects. However, lack of a corpus for these dialects is a constraint. Our model needs nothing more than the corpus text and a binary

⁹ <https://bit.ly/2K45V19>

CE	CH	TW	TWSM	SE
A	अ	552	242	165
Aa	आ	80	47	36
Ai	ऐ	9	7	2
B	ब	817	391	254
Bha	भ	288	114	79
Ca	च	437	201	149
Chha	छ	177	79	59
Da	ड	161	67	48
Dha	ढ	78	36	26
E	ए	21	11	9
Fa	फ	201	109	69
Ga	ग	553	254	162
Gha	घ	163	70	51
Ha	ह	297	151	112
I	इ	44	30	21
Ii	ई	8	6	5
Ja	ज	322	196	131
Jha	झ	105	44	26
Kha	ख	365	194	132
L	ल	272	121	87
Ma	म	578	305	194
Na	न	335	180	129
O	ओ	42	26	17
Qu	औ	13	7	5
Pa	प	657	310	226
Ra	र	251	125	89
Sa	स	815	391	272
Ta	त	338	175	119
T'ha	थ	65	32	24
Ta	ट	130	66	50
Tha	ठ	86	41	28
Thha	ड	132	57	38
Tiha	द	375	189	138
LI	ड	173	75	54
Lu	ळ	14	9	6
Ya	व	40	19	12
Ya	य	33	18	12
Total		9027	4395	3036

CE: hindi alphabet in english, CH: hindi alphabet, TW: total Awadhi words in the dictionary for a given alphabet, TWSM: total Awadhi words with single word Hindi meanings, SE: Hindi synsets enriched in IndoWordnet

Table 3: Indowordnet Enrichment statistics

mapping of dialect words to the mother language; today's technological armoury is such that if such text is collected in any digital form (textual/ visual/ audio /video) the model can work through it.

Although we achieved encouraging results there are certain shortcomings which if taken care of can make this model more reliable.

Another improvement can be incorporated by designing stemmers for a given dialect. Right now the inverse mappings from bilingual dictionaries contain mappings of 'word-to-word' form but 'phrases-to-words' and 'n-grams' can be considered further. Through 'phrases-to-word' mappings we can decrease the inversion loss percentage of 48.48. Also, sentences generated by our model use the forward probability of words in a sentence. For capturing the complete context, backward probability can give better results. Overall, the model has good potential for further growth. Each dialect of Hindi has its own geographical style and culture. Our future aim would be to generate sentences using RNN that highlight these cultural aspects.

6. Bibliographical References

- Arun Baby, Anju Leela Thomas, Nishanthi N L, and Hema A Murthy, "Resources for Indian languages", CBBLR workshop, International Conference on Text, Speech and Dialogue. Springer, 2016.
- Bhattacharyya P (2010) IndoWordNet de Melo G, Weikum G (2012) Constructing and utilizing WordNets using statistical methods. *Lang Resour Eval* 46(2):287-311
- B.L. Sukhwal, *Modern Political Geography of India*, Stosius Inc/Advent Books Division, ... In the Hindi heartland ... (1985)
- E. G. Caldarola and A. M. Rinaldi, "Improving the Visualization of WordNet Large Lexical Database through Semantic Tag Clouds," 2016 IEEE International Congress on Big Data (BigData Congress), San Francisco, CA, 2016, pp. 34-41.
- Gandhe, Ankur, Florian Metzke, and Ian Lane. "Neural network language models for low resource languages." *Fifteenth Annual Conference of the International Speech Communication Association*. 2014.
- Islam, Md Sanzidul, et al. "Sequence-to-sequence Bangla sentence generation with LSTM Recurrent Neural Networks." *Procedia Computer Science* 152 (2019): 51-58.
- Khodak, Mikhail, Andrej Risteski, Christiane Fellbaum, and Sanjeev Arora. "Extending and Improving Wordnet via Unsupervised Word Embeddings." arXiv preprint arXiv:1705.00217 (2017).
- Kishore, Papineni, et al. "BLEU: a method for automatic evaluation of machine translation." *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002.
- Martha Palmer, Rajesh Bhatt, Bhuvana Narasimhan, Owen Rambow, Dipti Misra Sharma, Fei Xia. Hindi Syntax: Annotating Dependency, Lexical Predicate-Argument Structure, and Phrase Structure. In the Proceedings of the 7th International Conference on Natural Language Processing, ICON-2009, Hyderabad, India, Dec 14-17, 2009.
- Papineni, Kishore & Roukos, Salim & Ward, Todd & Zhu, Wei Jing. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. 10.3115/1073083.1073135.
- P. Malik and A. S. Baghel, "An improvement in BLEU metric for English-Hindi machine translation evaluation," *2016 International Conference on Computing, Communication and Automation (ICCCA)*, Noida, 2016, pp. 331-336.
- Ritesh Kumar, Bornini Lahiri, Deepak Alok, Atul Kr. Ojha, Mayank Jain, Abdul Basit, and Yogesh Dawar. 2018. Automatic identification of closely-related Indian languages: Resources and experiments. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC).
- Riyaz Ahmad Bhat, Rajesh Bhatt, Annahita Farudi, Prescott Klassen, Bhuvana Narasimhan, Martha Palmer, Owen Rambow, Dipti Misra Sharma, Ashwini Vaidya, Sri Ramagurumurthy Vishnu, and Fei Xia. The Hindi/Urdu Treebank Project. In the Handbook of Linguistic Annotation (edited by Nancy Ide and James Pustejovsky), Springer Press
- Sutskever, Ilya, James Martens, and Geoffrey E. Hinton. "Generating text with recurrent neural networks." *Proceedings of the 28th international conference on machine learning (ICML-11)*. 2011
- Taghizadeh, Nasrin, and Hesham Faily. "Automatic wordnet development for low-resource languages using cross-lingual WSD." *Journal of Artificial Intelligence Research* 56 (2016): 61-87.
- Varun Gandhi (2018), <https://economictimes.indiatimes.com/blogs/et-commentary/preserving-indias-endangered-languages/>
- Yadav, Arun Kumar and Samir Kumar Borgohain. "Sentence generation from a bag of words using N-gram model." *2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies* (2014): 1771-1776.

English WordNet 2020: Improving and Extending a WordNet for English using an Open-Source Methodology

John P. McCrae¹, Alexandre Rademaker², Ewa Rudnicka³, Francis Bond⁴

¹ Insight Centre for Data Analytics, NUI Galway, john@mccr.ae

² IBM Research and FGV/EMAP, alexrad@br.ibm.com

³ Wroclaw University of Science and Technology, ewa.rudnicka@pwr.edu.pl

⁴ Nanyang Technological University, bond@ieee.org

Abstract

The Princeton WordNet, while one of the most widely used resources for NLP, has not been updated for a long time, and as such a new project English WordNet has arisen to continue the development of the model under an open-source paradigm. In this paper, we detail the second release of this resource entitled “English WordNet 2020”. The work has focused firstly, on the introduction of new synsets and senses and developing guidelines for this and secondly, on the integration of contributions from other projects. We present the changes in this edition, which total over 15,000 changes over the previous release.

Keywords: WordNet, lexicons, open source, lexicography, NLP

1. Introduction

English WordNet (McCrae et al., 2019) is a fork of Princeton WordNet (Fellbaum, 2010; Miller, 1995), which aims to further the development of a wordnet for English. Wordnets are one of the most widely used resources in natural language processing¹ and as the English language is not static, it is necessary to continually update the resource so that it remains relevant for these tasks. Wordnets group words into sets of synonyms, called *synsets*, and each *sense* of a word corresponds to its membership in one synset. These synsets are then organized in a graph containing relationships such as *hypernym/hyponym* (broader or narrower), *antonym* and many more relations. The English WordNet has taken an open-source policy for this and the resource is available on GitHub,² where anyone can contribute to its development. A first release of this resource was made in 2019, although this release had some limitations in terms of the changes it made, in particular no new synsets were created in that release. In this paper, we describe the 2020 release, which provides a more thorough revision of the resource, including new synsets from other resources including Colloquial WordNet (McCrae et al., 2017), enWordNet (Rudnicka et al., 2015) and Open Multilingual WordNet (Bond and Paik, 2012). We also discuss some of the major challenges that we have encountered during the development of this resource. In particular, as the resource has introduced a large number of new synsets, we have had to develop guidelines for the significance of synsets to be included in the resource. We have also looked into making clearer guidelines for making sense distinctions between two meanings of the same word, as this seems to be a significant challenge for those who build systems on top of WordNet. Finally, we look at some of the challenges that we wish to address in the next year of the development, of which the most pressing is the adjective hierarchy, which is less dense and contains many unclear sense distinctions, as well as issues related to

improving the procedure for development of the resource, in particular with the format and the issue of ensuring backwards compatibility with Princeton WordNet.

2. Development Methodology

2.1. Open Source Development

Add Relation	5	Change Relation	18
Definition	44	Example	8
Delete Synset	8	New Synset	30
Synset Duplicate	32	Synset Member	19
Synset Split	1	Enhancements	8
Contribution	3	Bug	9

Table 1: The number of issues by type addressed in this release

English WordNet is based on an open source methodology and as such anyone can contribute to the development of this resource. We have developed a methodology as described previously (McCrae et al., 2019), that relies on **issues** and **pull requests** in order to manage requests for changes. While, there have been relatively few pull requests made directly to the project (in fact only 3 in the last year), issues have proven to be an effective method by which requests can be logged. In total 161 issues were created asking for changes in the WordNet, that have been closed as part of this release. The number of each type of issue is given in Table 1 where they are categorized according to the following scheme:

Add Relation A relation should be added between two synsets;

Change Relation A relation is of the wrong type or has the wrong target;

Definition The definition of a synset should be updated;

Example The examples of a synset should be updated;

¹The original paper has over 13,000 citations

²<https://github.com/globalwordnet/english-wordnet>

Delete Synset A synset represents a concept that should be removed from WordNet. There are few reasons for this: the concept cannot be found in any other reference material and there is no corpus evidence for its members; the synset refers to a compositional meaning; the synset exists twice in the wordnet;

New Synset A synset covering a new concept is being proposed;

Synset Duplicate Two synsets are not possible to distinguish or refer to the same concept. This is fixed by either creating a new concept for all synsets or by deleting all but one of the duplicates;

Synset Split A synset refers to two distinct concepts and should be split into two new synsets;

Synset Member A word in a synset should be added or removed;

Enhancement A request for an improvement in the tooling around English WordNet or for a new kind of data;

Contribution Issues related to large external contributions (see Section 3.);

Bug A technical flaw that needs to be addressed in the data files.

Once these issues have been logged then a solution is proposed by one of the team members and a pull request is made, and then accepted. The process is designed to give high visibility to the changes proposed in the wordnet (which has helped to detect minor errors) and to provide tracking so that the discussion and implementation can be easily connected through Git. This means that the changes are all well-documented and as such could easily be taken up by other projects or included back into Princeton WordNet.

2.2. Guidelines for new synsets

As this version of English WordNet has introduced new synsets it has been necessary to formalize the guidelines for the introduction of new synsets. These guidelines attempt to formalise best practices from Princeton WordNet and other projects and they are based on the principle that new synsets should be added with some caution. In fact, they are much stricter than the current set of synsets that are derived from Princeton WordNet, thus if applied retroactively would lead to the removal of many existing synsets, and is not planned for the foreseeable future.

We have defined five basic criteria that a new synset is required to pass before being introduced into the wordnet: (1) Significance; (2) Non-compositionality; (3) Distinction; (4) Well-defined; and (5) Linked.

2.2.1. Significance

A concept in English WordNet should be significant, this means that it should be possible to easily find *at least 100 examples* of the usage of the word with this meaning. This

can be done by using a search interface such as Sketch Engine³ or other corpus search interface. For future releases, we aim to integrate corpora tools into the GitHub instance. In the case that a new sense of an existing word is being proposed, then it should be possible to propose collocates that occur with this sense of the word and these can be used to find and distinguish examples.

English WordNet is a dictionary not an encyclopedia. For this reason, it should not contain long lists of people, places, organizations, etc. Proper nouns are generally not expected to be included in the resource and many kinds of common nouns for narrow domains or geographical usage should not be included, examples of this would include elements of different cuisines around the world. As a rule of thumb, if there is a Wikipedia page for this concept it should not be in English WordNet.⁴ For future releases a more complete alignment of the resource and Wikipedia is planned based on previous works (De Melo and Weikum, 2009; McCrae, 2018) to address the introduction of synsets already well-described in Wikipedia.

2.2.2. Non-compositionality

One of the goals of English WordNet is to support annotation. If a word (or multiword expression) is already covered by English WordNet it should not be added.

For multiword expressions (MWE), this means that the meaning of the term should not be derivable from its components, e.g., “French Army” could be tagged with the synsets for “French” and “Army”; in contrast “operational system” refers not to a system that is operational, but it is a computer science term for the system that runs on every computer. Another case of MWE is the conventionalized ones. Conventionalization refers to the situation where a sequence of words that refer to a particular concept is commonly accepted in such a way that its constituents cannot easily be substituted for near-synonyms, because of some cultural or historical conventions (Farahmand et al., 2015). Consider the expression “geologic fault”. It is compositional but no one would consider substituting it with “geologic defect”. There are many types of MWE and a extensive literature about them (Sag et al., 2002), here we just want to emphasize that expressions that could have their parts annotated with senses already in the resource don’t need to be explicitly added.

For single words, the word should not be derived in a systematic manner, these include:

- Converting a verb to a noun or adjective by adding ‘-ing’ or ‘-ed’
- Converting an adjective to an adverb by adding ‘-ly’
- Productive prefixes such as ‘non-’, ‘un-’
- Systematic polysemy: e.g., using a part to refer to a whole, for example: “congress” meaning the “members of congress”

³<http://sketchengine.eu>

⁴There is no plan to apply this retroactively to existing synsets at the moment

2.2.3. Distinction

The concept should be distinct from other concepts in the WordNet and care should be taken to check relevant synonyms. For each word in the synset, the sense should thus be distinct as described above. This is best considered in terms of a substitution check, e.g., “happy” and “felicitous” are synonyms, `ewn-01052105-s` and the examples can be substituted, e.g., “a happy life”/“a felicitous outcome”. This does not mean that they can be substituted in every sense, e.g., “happy to help” but not *“felicitous to help”.

2.2.4. Well-defined

It should be possible to easily write a definition for this concept that is distinct from other concepts in English WordNet. A good definition consists of a *genus* and a *differentia*.

Genus The type of the thing, often the hypernym,

Differentia Something that makes this word unique

An example of a good definition is:

a piece of furniture having a smooth flat top that is usually supported by one or more vertical legs

Where a poor definition would be:

a piece of furniture

used for eating

In addition an example should be provided with a link to a website where the example is used as follows:

```
<Synset id="ewn-...">
  ...
  <Example dc:source=
"https://en.wikipedia.org/wiki/Example.com">
  The example domains have one subdomain
  name defined in the Domain Name System
  </Example>
  ...
</Synset>
```

2.2.5. Linked

The synset should be possible to link into the graph, more specifically:

Nouns A hypernym must be identified

Verbs A hypernym, entailment, cause or antonym must be identifier. Verbs should also have at least one subcategorization frame.

Adjectives They should be marked as similar to a non-satellite adjective (in which case they are satellites) **or** antonyms of a non-satellite adjective **or** hypernyms of an adjective

Adverbs No clear guidelines but at least one link should be proposed. Ideally a link for the corresponding adjective via derivation relation.

The more links that can be provided the better a synset is.

2.2.6. Sense keys and lexicographer files

Two key design aspects that are derived from Princeton WordNet are the use of lexicographer files and sense keys, however the changes in the development procedure for English WordNet (as opposed to Princeton WordNet) have made into necessary to update how these elements are used. English WordNet is divided into a number of source files that correspond to the original lexicographer files in WordNet, but are now in XML. New synsets proposed from issues should be assigned to one of these lexicographer files as they are created. For contributed resources (see below), we merged them into the original resource according to the hypernym.

Sense keys were a mechanism that provided stability between releases of WordNet, and sense keys were (mostly) stable identifiers between different versions of Princeton WordNet. Instead, English WordNet has adopted the CILI interlingual index (Vossen et al., 2016; Bond et al., 2016) as the principal method of providing cross-version stability. Moreover, for new senses the calculation of stable sense identifiers is complicated as the Princeton WordNet formula relied on information in lexicographer files that is no longer present. Initial proposals were just to jettison sense keys, however community feedback has encouraged the creation of new methodology for assigning sense keys.⁵ In addition, we now also track the changes of sense keys, caused for example changes in the spelling of a lemma or if a sense has been moved across lexicographer files.

2.2.7. Sense distinctions

One particular issue that has been common in the reported set of issues is the issue of sense distinction. WordNet has been criticized (Palmer et al., 2004; Snow et al., 2007) for a long time for issues related to its sense granularity. As such, there have been many issues claiming that synsets are duplicated as the meanings are quite hard to distinguish. In order to simplify these decisions, we have developed a few key principles that help us in distinguishing senses.⁶

Ontological Typing Two synsets that have difficult to distinguish senses may typically occur in different parts of the WordNet graph. This can often make the distinction clearer than the definitions as the two synsets refer to ontological distinct aspects. For example, for ‘rock’,⁷ the definitions were not clear however the structure clearly gave away that the two senses referred to ‘rock’ as a material and ‘rock’ as a physical object, that is the first sense was uncountable and the second countable.

Collocations Following methods in word sense induction (Klapaftis and Manandhar, 2008; Denkowski, 2009), one clear rule for distinguishing two senses is the existence of collocations that cannot be applied to both senses. We aim in the future to extend this basic principal with some quantitative scoring function that can help us in distinguishing senses based on corpus information. For example, ‘rock’ collocated with

⁵Issue #157.

⁶Track with Issue #243

⁷Issue #135

‘concert’ suggests a very different sense to a collocation with ‘metamorphic’.

Other dictionaries The final method we use for deciding whether to make a sense distinction is to look at other dictionaries. In cases, where a very subtle distinction is being discussed often comparison with other dictionaries can help to decide these issues.

3. Integration of Existing Resources

3.1. Colloquial WordNet

POS	Lemmas	Synsets
Noun	196	195
Verb	75	79
Adjective	34	36
Adverb	5	5
Total	310	315

Table 2: New synsets and lemmas introduced by Colloquial WordNet by part of speech

Colloquial WordNet (McCrae et al., 2017) was a resource developed to extend wordnet with recent slang terms. The resource included a number of changes that would not be in line with the existing wordnet, although may be later included as these features are added to the mainstream of wordnet. These include the marking of non-referential expressions (such as “ah!”, or “haha”), the sense linking from a multiword expression to the senses of its individual words and the mark of words as loanwords from other languages. Once these had been removed the resource was integrated, which is relatively simple as the Colloquial WordNet uses the same format as English WordNet. However the new synsets introduced by this wordnet were given 8-figure numeric codes much like in the existing wordnet. As these cannot be based on the offset in a file, instead they were assigned based on the original identifiers with a code starting 90 or 91. For example, ‘adulting’⁸ is code `ewn-900004011-n`.

3.2. Open Multilingual WordNet

The Open Multilingual WordNet (Bond and Paik, 2012; Bond and Foster, 2013) project has also introduced new synsets and made changes related to the English WordNet. We are in the process of integrating these changes, one of the most major changes is the rewriting of definitions so as to ensure uniqueness. This change affects 1,673 synsets and most of these changes directly improve the definitions as given, for example, ‘Thai’ was previously defined as ‘a branch of the Tai languages’ and is now defined as ‘a branch of the Tai languages, spoken in central Thailand, centered in Bangkok’.

In addition, there are a large number of changes that introduce new synsets, mostly to cover concepts that are not already in the wordnet. We are currently in the process of identifying these changes and integrating those that meet the guidelines for new synsets.

⁸“acting like an adult”

3.3. enWordNet

The plWordNet team at Wrocław University of Science and Technology has also developed a number of extensions of the English WordNet (Zaško-Zielińska and Piasecki, 2018; Dziob and Piasecki, 2018; Janz et al., 2017) to cover concepts not currently covered in English WordNet. We are integrating these changes into our format. In total, the enWordNet (as of version 4.0) has proposed 7,656 new synsets, however our analysis quickly deduced that many of them consist of concepts that are easily found in Wikipedia and are defined by sections of text copied from Wikipedia. We automatically reduced the set of proposed changes to 2,084 synsets by applying the guidelines in Section 2.2., in particular by looking for lemmas that match existing Wikipedia page titles. We then conducted a manual review of this, we found that 1,843 out of 2,084 (88.4%) synsets were of acceptable quality to be introduced in English WordNet. This represents a large part of the changes that have been made in the 2020 release.

4. Open Challenges

4.1. Satellite Adjectives

As previously discussed, sense distinctions have been an important difficulty in the development of the resource. For adjectives, most of these issues have not yet been solved as the structure of adjectives in WordNet is currently quite suboptimal. In particular, English WordNet distinguishes between two kinds of adjectives: ‘head’ adjectives and satellite adjectives. Head adjectives should have an antonym relation to another head adjectives, which satellite adjectives should be marked as similar to a head adjective; this is called the ‘dumbbell’ model. The distinction is made at the part-of-speech level in the resource, although no other part-of-speech catalogue or dictionary to our knowledge makes the distinction this way.⁹ This means that there is often fewer links to other synsets and also shorter definitions; in fact adjectives typically have 1.44 synset links against a general average of 2.43. The plan for a future version, is to revamp the adjective so that they follow a more conventional classification such as that proposed by (McCrae et al., 2014), where the formal categories are:

Intersective These refer to properties that the adjective indicates the presence of. The most significant group of these are pertainyms, which mean that a concept is of or pertaining to a noun, e.g., “French” pertaining to “France”. The existing pertainym relation marks many of these but can be expanded.

Gradable These adjectives refer to the value of a property on some scale, for example ‘hot’ is on a scale of ‘temperature’, a new property relating adjectives to their scales will be introduced and this will replace the ‘dumbbell’ model.

Operator This group will capture that final set of adjectives that have a meaning that modifies the meaning

⁹This is even though more widely-accepted distinction such as postpositive adjectives are distinguished at the sub-part-of-speech level

of the noun, such as ‘former’. We will look into new properties that could be introduced to help with connecting these concepts in the WordNet graph.

4.2. Format

The English WordNet is currently published under the GWA XML format, however there have been a number of issues related to this, most principally that the format is quite verbose as is typical for XML. Moreover, we have found that some aspects of the data contained in the original Princeton WordNet are not possible to represent in this format.

There are two proposals for moving from the XML format:¹⁰ the first is to stick with the GWA model but use a less verbose serialization (namely YAML) and reduce the amount of information represented in the dictionary files. The second option is to adopt the model presented in Muniz et al. (2018) and being investigated by the OpenWordnet-PT project (Paiva et al., 2012),¹¹ as there is a large amount of work already carried out with this model, however it is a non-standard serialization and due to its brevity it can be difficult to understand for those not used to it.

In addition, there are a number of problems related to the representation of existing data from Princeton WordNet, these include the morphosemantic relations that are provided in a second stand-off file¹² and these can be easily included in the main resource simply by extending the set of relations that are available in the WordNet. The next issue is related to adjective position,¹³ which was not captured in the previous release and cannot be encoded in the XML format as the part-of-speech categories are a closed group. We have added this as a new attribute, `adjposition`, on the `LexicalEntry` tags in the resource. Finally, there were some verb example sentences,¹⁴ that were not being captured. This was after it was discovered that the previous release was using the wrong file to generate the syntactic behaviour of the entries. As such, this will be added as new examples on the corresponding synset with an extra attribute to say that they were generated by sentence templates.

4.3. Backwards Compatibility

One of the key goals of the model is to ensure that there is backwards compatibility with previous Princeton WordNet releases, and as such, although the project has moved to the XML format entirely, we still make releases in the previous WNDB format. This leads to a number of issues, most notably that synset identifiers are based on file offsets in this format. In particular, as we do not wish to recalculate the identifiers used in the XML files at every release the identifiers in the WNDB release will not correspond to those in the XML. This is further exacerbated by the introduction of new synsets, whose identifier is set to be high enough

that it cannot correspond to a byte offset in the file. In addition, there have been a number of issues related to sense identifiers that have been improved in this release to provide more continuity for users of English WordNet in the WNDB format.

Finally, as the license of WordNet is unique to Princeton WordNet, we are moving to use a Creative Commons Attribution license to protect the changes made on top of Princeton WordNet. As the underlying resource (Princeton WordNet) has its own bespoke license, it is necessary to reproduce both licenses when deriving resources from English WordNet.¹⁵

4.4. Distributed model

It is not clear how domain-specific or goal-specific wordnets (such as the Colloquial WordNet) should be incorporated or linked to the English Wordnet. Regarding the data format, a linked-open data format such as RDF could help us in the definition of global identifiers (URI) that could help on the link of entities in different resources. But this is part of the problem, the maintainance of the links and the track of changes on these resources can be far from trivial. On the other way, incorporating domain-specific or goal-specific wordnets into English Wordnet would make the resource maintainance even harder with increasing difficult on the definition of guidelines such as the ones explained above.

5. Changes in 2020 Release

The total number of changes are detailed in Table 3, and as can be seen the largest number of changes are firstly to do with the definitions. This is due to the contribution of many new definitions from Colloquial WordNet and enWordNet and secondly, to do with the many changes proposed by the Open Multilingual WordNet project. Secondly, we see a large number of new lemmas and synsets proposed by both Colloquial WordNet and enWordNet, representing the largest number of changes. As many of these are single nouns whose lemma does not already occur in WordNet, the majority of the changes result in one new synset, one new lemma, one new sense and two more synset relations (typically a hypernym and a hyponym). While much effort has gone into the directly reported issues, most of these result in only small changes to the structure of the wordnet. We also see a lot of changes in the senses, this is primarily due to the change in the representation of adjective categories (e.g., postpositivity) as discussed above.

6. Conclusion

English WordNet is continuing to grow and meet the annual release schedule, to ensure that an up-to-date and accurate WordNet is available for the many users of WordNet in natural language processing. The open-source methodology that has been adopted has been generally successful so far and has provided impetus for the development of clear guidelines that are easy-to-follow. In this paper, we have discussed guidelines for new synsets and senses and detailed some of the open challenges that we are looking into, including the structuring of adjectives.

¹⁰Discussion is to be found at [Issue #31](#)

¹¹This format is called [Mill](#)

¹²[Issue #132](#)

¹³[Issue #180](#)

¹⁴[Issue #245](#) definition in [WNDB](#)

¹⁵[Issue #144](#)

	Princeton WordNet 3.1	English WordNet 2019	English WordNet 2020	Changed
Synsets	117,791	117,791	120,054	
Lemma	159,015	159,789	163,079	
Senses	207,272	208,353	211,864	
Synset Relations	285,668	285,666	291,299	
Sense Relations	92,535	92,535	92,526	
Definitions	117,791	117,791	120,059	1,587
Examples	47,539	48,419	49,675	151

Table 3: Comparative size of Princeton WordNet 3.1 and English WordNet 2019 and 2020

Acknowledgements

This work is supported by the EU H2020 programme under grant agreements 731015 (ELEXIS - European Lexical Infrastructure). John McCrae is also supported by a research grant from Science Foundation Ireland, co-funded by the European Regional Development Fund, for the Insight Centre under Grant Number SFI/12/RC/2289. Ewa Rudnicka is supported by the CLARIN-PL project, which is part of the investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education.

7. Bibliographical References

- Bond, F. and Foster, R. (2013). Linking and extending an open multilingual wordnet. Sofia.
- Bond, F. and Paik, K. (2012). A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, Matsue. 64–71.
- Bond, F., Vossen, P., McCrae, J. P., and Fellbaum, C. (2016). CILI: the Collaborative Interlingual Index. In *Proceedings of the Global WordNet Conference 2016*.
- De Melo, G. and Weikum, G. (2009). Towards a universal wordnet by learning from combined evidence. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 513–522.
- Denkowski, M. (2009). A survey of techniques for unsupervised word sense induction. *Language & Statistics II Literature Review*, pages 1–18.
- Dziob, A. and Piasecki, M. (2018). Implementation of the verb model in plWordNet 4.0. In *Proceedings of the 9th Global Wordnet Conference*, pages 114–123.
- Farahmand, M., Smith, A., and Nivre, J. (2015). A multiword expression data set: Annotating non-compositionality and conventionalization for English noun compounds. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 29–33.
- Fellbaum, C. (2010). WordNet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.
- Janz, A., Kocoń, J., Piasecki, M., and Zaśko-Zielińska, M. (2017). plWordNet as a basis for large emotive lexicons of Polish. *Proceedings of Human Language Technologies as a Challenge for Computer Science and Linguistics Poznań: Fundacja Uniwersytetu im. Adama Mickiewicza w Poznaniu*, pages 189–193.
- Klapaftis, I. P. and Manandhar, S. (2008). Word sense induction using graphs of collocations. In *ECAI*, pages 298–302.
- McCrae, J. P., Unger, C., Quattri, F., and Cimiano, P. (2014). Modelling the Semantics of Adjectives in the Ontology-Lexicon Interface. In *Proceedings of 4th Workshop on Cognitive Aspects of the Lexicon*.
- McCrae, J. P., Wood, I., and Hicks, A. (2017). The Colloquial WordNet: Extending Princeton WordNet with Neologisms. In *Proceedings of the First Conference on Language, Data and Knowledge (LDK2017)*.
- McCrae, J. P., Rademaker, A., Bond, F., Rudnicka, E., and Fellbaum, C. (2019). English WordNet 2019 – An Open-Source WordNet for English. In *Proceedings of the 10th Global WordNet Conference GWC 2019*.
- McCrae, J. P. (2018). Mapping WordNet Instances to Wikipedia. In *Proceedings of the 9th Global WordNet Conference*.
- Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.
- Muniz, H., Chalub, F., Rademaker, A., and de Paiva, V. (2018). Extending wordnet to geological times. In *Global Wordnet Conference 2018*, Singapore, January.
- Paiva, V. d., Rademaker, A., and Melo, G. d. (2012). OpenWordNet-PT: An open Brazilian WordNet for reasoning. Technical report, COLING 2012.
- Palmer, M., Babko-Malaya, O., and Dang, H. T. (2004). Different sense granularities for different applications. In *Proceedings of the 2nd International Workshop on Scalable Natural Language Understanding (ScaNaLU 2004) at HLT-NAACL 2004*, pages 49–56.
- Rudnicka, E., Witkowski, W., and Kaliński, M. (2015). Towards the Methodology for Extending Princeton WordNet. *Cognitive Studies*, 15(15):335–351.
- Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for nlp. In *International conference on intelligent text processing and computational linguistics*, pages 1–15. Springer.
- Snow, R., Prakash, S., Jurafsky, D., and Ng, A. Y. (2007). Learning to merge word senses. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (emnlp-conll)*, pages 1005–1014.
- Vossen, P., Bond, F., and McCrae, J. P. (2016). Toward a truly multilingual Global Wordnet Grid. In *Proceedings of the Global WordNet Conference 2016*.
- Zaśko-Zielińska, M. and Piasecki, M. (2018). Towards emotive annotation in plWordNet 4.0. In *Proceedings of the 9th Global Wordnet Conference, Singapore*, pages 154–163.

Exploring the Enrichment of Basque WordNet with a Sentiment Lexicon

Jon Alkorta, Itziar Gonzalez-Dios

Ixa Group, University of the Basque Country (UPV/EHU)

Manuel Lardizabal Ibilbidea, 1, 20018 Donostia

{jon.alkorta,itziar.gonzalezd}@ehu.eus

Abstract

Wordnets are lexical databases where the semantic relations of words and concepts are established. These resources are useful for many NLP tasks, such as automatic text classification, word-sense disambiguation or machine translation. In comparison with other wordnets, the Basque version is smaller and some PoS are underrepresented or missing e.g. adjectives and adverbs. In this work, we explore a novel approach to enrich the Basque WordNet, focusing on the adjectives. We want to prove the use and effectiveness of sentiment lexicons to enrich the resource without the need of starting from scratch. Using as complementary resources, one dictionary and the sentiment valences of the words, we check if the word of the lexicon matches with the meaning of the synset, and if it matches we add the word as variant to the Basque WordNet. Following this methodology, we describe the most frequent adjectives with positive and negative valence, the matches and the possible solutions for the non-matches.

Keywords: Basque_WordNet, sentiment_lexicon, resource enrichment

1. Introduction and Background

Creating and maintaining language resources is an expensive and costly task. Moreover, in the case of the languages with recent standardisation processes, the update and constant redesign of the resources is mandatory. In the case of Basque, a language whose standardisation process officially began in 1968, the maintenance and updating lexicosemantic resources such as the Basque Wordnet (BWN) or *Euskal Wordnet* (Pociello et al., 2011) requires a big lexicographic effort (Aldezabal et al., 2018).

BWN¹ is a version in Basque language of WordNet (Miller, 1995; Fellbaum, 1998). It was created following the expand approach, but special care was taken for cultural concepts and lexicalization issues. It was developed together with the EuSemCor corpus (Agirre et al., 2006). In the latest distribution (2016), BWN had 30 263 synsets, 40 420 variants for nouns, 9 469 for verbs and 148 for adjectives. There were no variants for adverbs. As far as the size of BWN is concerned, its size is limited in comparison with wordnets in other languages. As updating it from scratch is costly, in this paper we explore multimodal approaches to add new variants, namely based on the sentiment lexicon for Basque *SentiTegi* (Alkorta et al., 2018)

Regarding wordnets and sentiment lexicons, wordnets are usually complemented with polarity and sentiment information. They have been created above all for sentiment analysis and opinion mining. For example, in SentiWordNet 3.0 (Baccianella et al., 2010) each synset is tagged with the notions positivity, negativity, and neutrality and associated to three numerical scores to indicate how positive, negative, and objective/neutral the variants are. Based on SentiWordNet, SentiWord (Gatti et al., 2015) profits from prior built polarity lexica in order to achieve higher precision and coverage. Finally, in WordNet-feelings (Siddharthan et al., 2018) synsets are classified in nine broad feeling categories

and as a feeling based on a depth linguistic study of human feelings. This resource is complementary to SentiWordNet. All these resources have been created based on the English WordNet.

In this paper, however, we want to explore the opposite option: the possibility of using the sentiment lexicon *SentiTegi* to increase the size of BWN. We want to focus on the adjectives, whose coverage is limited in BWN. Exactly, as case study we will analyze the most frequent adjectives with positive and negative valence in *SentiTegi*, is a manually-created sentiment lexicon for Basque. The aim is to explore if *SentiTegi*, and to a certain extend sentiment lexicons, can be used as a source to enrich BWN and wordnets. Moreover, we want to examine which linguistic issues arise when comparing both resources.

This paper is structured as follows: in Section 2 we describe the characteristics of the sentiment lexicon *SentiTegi*, which is the source for the enrichment of the BWN. In Section 3, we explain the casuistry regarding the match between meanings of synsets and the sentiment valence of the words by some examples of the enrichment. Finally, in Section 4, we conclude the work and enumerate the future works.

2. *SentiTegi*, the Basque Sentiment Lexicon

SentiTegi (Alkorta et al., 2018) is a manually-created sentiment lexicon for Basque. It is a part of the Basque version of the sentiment classifier called SO-CAL (Taboada et al., 2011). The SO-CAL sentiment classifier is a lexicon-based sentiment classifier. The words in lexicon have a sentiment valence² that determines if the words make a positive or negative evaluations or judgements. The sentiment valence of the words is numerical and the numbers rank from -5 to $+5$.

²Sentiment valence and semantic orientation are used to determine the subjectivity of words. Semantic orientation is a signal (+ or -) that indicate if the word makes a positive (or good) or negative (or bad) evaluation. In contrast, the sentiment valence indicates the intensity of the evaluation with numbers in addition to type of evaluation (good or bad evaluation).

¹The Basque Wordnet is available in <https://adimen.si.ehu.es/cgi-bin/wei/public/wei.consult.perl> and in the Open Multilingual wordnet <http://compling.hss.ntu.edu.sg/omw/>

- (1) *Bikain* “excellent” (+5)
- (2) *Eskas* “insufficient” (−1)
- (3) *Txar* “bad” (−3)

Examples (1), (2) and (3) indicate the scale of the words in the sentiment lexicon regarding their subjective evaluation. In Example (1), the word *bikain* “excellent” expresses the most intense positive evaluation and, consequently, its sentiment valence is +5. On the other hand, in Example (2), the word *eskas* “insufficient” makes a negative evaluation, therefore the sign is negative (−). In contrast with Example (1), its intensity is lower and, for that reason, its sentiment valence is (−1). Finally, in Example (3), the word *txar* “bad” makes a negative evaluation and it has a medium intensity. Therefore, its sentiment valence is (−3).

The sentiment lexicon in Basque has been created by translating the sentiment lexicon in Spanish and enriching with the sentiment lexicon in English of different language versions of the SO-CAL tool. First, the sentiment lexicon from Spanish (Brooke et al., 2009) was translated into Basque with the *Elhuyar* (Elhuyar Hizkuntza Zerbitzuak, 2013) and *Zehazki* (Sarasola, 2005) dictionaries. In the second step, the Basque translations have been grouped with different source in Spanish lexicon. For example, the Spanish words *amago* “feint” (−1) and *cicatriz* “scar” (−2) have been translated into Basque as *seinale* “signal” and they have been grouped in the Basque word *seinale* “signal”. In addition, the Basque translations have inherited the sentiment valence from Spanish words. In the case of the words that had various sources, the most adequate for Basque has been chosen. Consequently, the sentiment valence (−1) has been chosen for *seinale* “signal”, based on the Spanish *amago* “feint”.

In order to choose the sentiment valence of the Basque words, the context where the words appear in the *Basque Opinion Corpus* (Alkorta et al., 2017) has been taken into account. Then, the Basque translations that are not entries of the *Elhuyar* (Elhuyar Hizkuntza Zerbitzuak, 2013) and *Zehazki* (Sarasola, 2005) dictionaries or do not appear in the *Basque Opinion Corpus* (Alkorta et al., 2017) have been removed. As a consequence of that, e.g. the word *atrofiatu* “atrophy” (−1) has been removed from the lexicon. After this step, the size of the sentiment lexicon has been reduced from 8,140 words to 1,237 words. Finally, the translated sentiment lexicon has been enriched with the English sentiment lexicon (Taboada et al., 2011). The sentiment valences of the Basque words and their equivalents in English have been compared: in some cases the sentiment valence has been changed and in other cases, the sentiment valence of the Basque word has been kept.

Table 1 shows the characteristics of the sentiment lexicon in Basque. The first version of the sentiment lexicon has been created following the first two steps (in other words, until grouping the Basque translations and choosing their sentiment valence). In contrast, the second version has been created following all the steps mentioned before. Among the first and second version, the size of the lexicon has been reduced from 8,140 words to 1,237 words due to the constraints of dictionaries and domains of the corpus. However, in both corpora, nouns and adjectives are the most

Part-Of-Speech	V1.0		V2.0	
	Words	%	Words	%
Nouns	2,282	28.06	461	37.27
Adjectives	3,162	38.85	446	36.05
Adverbs	652	7.98	54	4.36
Verbs	1,657	20.36	276	22.32
Intensifiers	387	4.75		
Total	8,140	100	1,237	100

Table 1: The characteristic of two versions of the sentiment lexicon *SentiTegi*

common grammatical category.

3. Methodology and Casuistry

In this section, we explain the methodology and the casuistry in the enrichment of the BWN with *SentiTegi*. In order to select the sample for the analysis, we have extracted the most frequent adjectives in *Basque Opinion Corpus* (Alkorta et al., 2017) with *AnalHitza* (Otegi et al., 2017), a tool that extracts basic linguistic information from texts and corpora. We have also filtered the adjectives taking into account their valence. The list we have created has the following information the Basque adjective, frequencies, valences and the respective English equivalents. For example, the Basque adjective *handi*, has a frequency of 101, its valence is +1, and its English equivalent is “big”.

Once having the list with frequencies, valences and the respective English equivalents, we have looked up the English in the word in the MCR. Taking also as a reference the *Euskaltzaindiaren Hiztegia* (Euskaltzaindia, 2016) the dictionary of the Academy of Basque Language that includes definitions, we have checked if the meaning of Basque word corresponds to the meaning of the synsets that contained the English variants. In addition to that, we also use the sentiment valence of the Basque adjectives to determine if the variant corresponds to the synsets.

In the following subsections, we explain the casuistry we have found for the adjectives by means of some examples.

3.1. Adjectives with positive semantic orientation

In Table 2, we show the most frequent positive adjectives of the sentiment lexicon *SentiTegi* in the *Basque Opinion Corpus*. Following, we present the analysis for these cases.

Basque	Instances	Sentiment valence	English	Sentiment valence
<i>Handi</i>	101	+1	<i>Big</i>	+1
<i>Berri</i>	57	+2	<i>New</i>	+2

Table 2: Two positive words with their sentiment valence taken from *SentiTegi*

As far as *handi* “big” is concerned, its sentiment valence is (+1). If we want to enrich BWN with this word of the sentiment lexicon, the sentiment valence of the Basque word and the the meaning of synset need to agree. According to

the sentiment lexicon, *handi* “big” means something positive because the sentiment valence is (+1). Moreover, the variant “big” appears in several synsets.

Synset	Meaning	Match
ili-30-01382086-a	above average in size or number or quantity or magnitude or extent	Yes
ili-30-01276872-a	significant	Yes
ili-30-01510444-a	very intense	Yes
ili-30-01453084-a	loud and firm	Yes
ili-30-00579622-a	conspicuous in position or importance	No
ili-30-02402439-a	prodigious	Yes
ili-30-01890752-a	exhibiting self-importance	No
ili-30-01890187-a	feeling self-importance	No
ili-30-01488616-a	(of animals) fully developed	No
ili-30-01191780-a	marked by intense physical force	Yes
ili-30-01114658-a	generous and understanding and tolerant	Yes
ili-30-01111418-a	given or giving freely	Yes
ili-30-00173391-a	in an advanced stage of pregnancy	No

Table 3: Synsets including the variant “big” and its matches with the meaning and valence of *handi*

In Table 3.1., we list the synsets and meanings related to the word “big”. Eight of them (ili-30-01382086-a, ili-30-01276872-a, ili-30-01510444-a, ili-30-01453084-a, ili-30-02402439-a, ili-30-01191780-a, ili-30-01114658-a, ili-30-01111418-a) match with sentiment valence and meaning of the word *handi* “big”. The match happens with the synsets associated to the intensity in different ways (physical or psychic) but not with self-importance. In these last cases (ili-30-00579622-a, ili-30-01890752-a, ili-30-01890187-a, ili-30-01488616-a, ili-30-00173391-a), “big” makes negative evaluation and it does not match with the sentiment valence of the word *handi* (+1). To express those physiological features, there is another word in Basque: *handinahi* “arrogant” and this word should be included in those synsets. Morphologically, *handinahi* is a compound and includes *handi* “big”, but it is an independent word. However, this lead us to think that derivative words and compounds can also play a role towards an automatic candidate proposal for non-matching synsets.

Regarding the word *berri* “new”, presented in Table 3.1., it matches with all the synsets (ili-30-01640850-a, ili-30-01687167-a, ili-30-00937186-a, ili-30-00128733-a, ili-30-02070491-a, ili-30-02584699-a, ili-30-01687965-a and ili-30-00818008-a) and their meanings. In addition, the novelty means something positive or good and it goes in line with the sentiment valence (+2) of the word. However, in some meanings like ili-30-00024996-a, the novelty could

Synset	Meaning	Match
ili-30-01640850-a	not of long duration; having just (or relatively recently) come into being or been made or acquired or discovered	Yes
ili-30-01687167-a	original and of a kind not seen before	Yes
ili-30-00937186-a	lacking training or experience	Yes
ili-30-00128733-a	having no previous example or precedent or parallel	Yes
ili-30-02070491-a	other than the former one(s); different	Yes
ili-30-02584699-a	unaffected by use or exposure	Yes
ili-30-01687965-a	(of a new kind or fashion) gratuitously new	Yes
ili-30-00818008-a	(of crops) harvested at an early stage of development; before complete maturity	Yes
ili-30-00024996-a	unfamiliar	Yes(?)

Table 4: Synsets related to word “new” in the BWN

be positive or negative according to the context.

3.2. Adjectives with negative semantic orientation

In this subsection we analyse the most frequent adjectives that have negative connotation. We present these synsets in Table 5.

Basque	Instances	Sentiment valence	English	Sentiment valence
<i>Politiko</i>	33	-1	<i>Political</i>	-1
<i>Txiki</i>	30	-1	<i>Little</i>	-1

Table 5: Three negative words with their sentiment valence taken from *SentiTegi* Alkorta et al. (2018)

The examples in Table 5 show a different casuistry regarding the match with the meaning of synsets.

In the case of the sentiment word *politiko* “political”, its meaning matches with three possible meanings (ili-30-01814385-a, ili-30-02857407-a and ili-30-02857587-a). But, when it comes to semantic orientation, the meanings of “political” in the English *WordNet* are neutral while in the case of the word *politiko* “political” is (-1). Therefore, there is a disagreement from the point of view of sentiment analysis. This suggests us that another synset may be necessary for this variant.

Finally, in the case of the sentiment word *txiki* “little”, there are two cases. The word matches with some synsets (ili-

Synset	Meaning	Match
ili-30-01814385-a	involving or characteristic of politics or parties or politicians	Yes(*)
ili-30-02857407-a	of or relating to your views about social relationships involving authority or power	Yes(*)
ili-30-02857587-a	of or relating to the profession of governing	Yes(*)

Table 6: Synsets related to word “political”

Synset	Meaning	Match
ili-30-01391351-a	limited or below average in number or quantity or magnitude or extent	Yes
ili-30-01554510-a	(quantifier used with mass nouns) small in quantity or degree; not much or almost none or (with ‘a’) at least some	Yes
ili-30-01649031-a	(of children and animals) young, immature	Yes
ili-30-01280908-a	(informal) small and of little importance	Yes
ili-30-01455732-a	(of a voice) faint	No
ili-30-02386612-a	ow in stature; not tall	Yes
ili-30-01467534-a	lowercase	No
ili-30-00855670-a	small in a way that arouses feelings (of tenderness or its opposite depending on the context)	Yes

Table 7: Synsets related to word “little” in the BWN

30-01391351-a, ili-30-01554510-a, ili-30-01649031-a, ili-30-01280908-a, ili-30-02386612-a and ili-30-00855670-a) but, in other cases, there is no match. For the synset ili-30-01455732-a, the word *baxu* “low” is more suitable than *txiki* and for the synset ili-30-01467534-a, the word *xeh* “minuscule” is more appropriate. So, in these cases, the variants for the concepts should be added from another resource.

4. Conclusion and Future Work

In this work we have explored a method to enrich the BWN using *SentiTegi*, the sentiment lexicon in Basque. *SentiTegi* contains words with semantic information (sentiment va-

lences, in this case) which are useful in the enrichment of the BWN with the help of a dictionary. In fact, in addition to the match of the definition, if the evaluation (positive or negative) of the meaning of synsets matches with the evaluation of the word (positive or negative sentiment valence), the word of the sentiment lexicon is valid for the BWN. This proves that *SentiTegi* and our methodology are good starting points for the enrichment of the BWN. However, we have found some cases where the direct addition of the word to BWN is doubtful. This leads us to think that criteria still need to be analysed and revised.

In future work, we would like to apply the Appraisal Theory (Martin and White, 2003) to this process of enrichment of the BWN. The Appraisal Theory is useful to categorize the type of subjectivity of words with sentiment valence. Indeed, not all the words with sentiment valence express the same sentiment. Some of them express opinions (for example, “hate”) and others express sentiments (for instance, “happy”). The annotation of sentiment words with this theory would help to identify better the synsets that would match with them.

5. Acknowledgements

This work has been partially funded by the the project DeepReading (RTI2018-096846-B-C21) supported by the Ministry of Science, Innovation and Universities of the Spanish Government, Ixa Group-consolidated group type A by the Basque Government (IT1343-19) and BigKnowledge – *Ayudas Fundación BBVA a Equipos de Investigación Científica 2018*.

6. Bibliographical References

- Agirre, E., Aldezabal, I., Etxeberria, J., Izagirre, E., Mendizabal, K., Pociello, E., and Quintian, M. (2006). A methodology for the joint development of the basque wordnet and semcor. In *LREC*, pages 23–28.
- Aldezabal, I., Artola, X., de Ilarraza, A. D., Gonzalez-Dios, I., Labaka, G., Rigau, G., and Urizar, R. (2018). Basque e-lexicographic resources: linguistic basis, development, and future perspectives. In *Workshop on eLexicography: Between Digital Humanities and Artificial Intelligence*.
- Alkorta, J., Gojenola, K., Iruskietia, M., and Taboada, M. (2017). Using lexical level information in discourse structures for basque sentiment analysis. In *Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms*, pages 39–47.
- Alkorta, J., Gojenola, K., and Iruskietia, M. (2018). Senti-tegi: Semi-manually created semantic oriented basque lexicon for sentiment analysis. *Computación y Sistemas*, 22(4).
- Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Senti-WordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Lrec*, volume 10, pages 2200–2204.
- Brooke, J., Tofiloski, M., and Taboada, M. (2009). Cross-linguistic sentiment analysis: From english to spanish. In *Proceedings of the international conference RANLP-2009*, pages 50–54.
- Elhuyar Hizkuntza Zerbitzuak. (2013). *Elhuyar hiztegia: euskara-gaztelania, castellano-vasco*. Elhuyar.

- Euskaltzaindia. (2016). *Euskaltzaindiaren Hiztegia*. Euskaltzaindia.
- Christiane Fellbaum, editor. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.
- Gatti, L., Guerini, M., and Turchi, M. (2015). SentiWords: Deriving a High Precision and High Coverage Lexicon for Sentiment Analysis. *IEEE Transactions on Affective Computing*, 7(4):409–421.
- Martin, J. R. and White, P. R. (2003). *The language of evaluation*, volume 2. Springer.
- Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.
- Otegi, A., Imaz, O., Díaz de Ilarraza Sánchez, A., Iruskieta Quintian, M., and Uria Garin, L. (2017). Analhitza: a tool to extract linguistic information from large corpora in humanities research.
- Pociello, E., Agirre, E., and Aldezabal, I. (2011). Methodology and Construction of the Basque WordNet. *Language resources and evaluation*, 45(2):121–142.
- Sarasola, I. (2005). *Zehazki: gaztelania-euskara hiztegia*. Alberdania.
- Siddharthan, A., Cherbuin, N., Eslinger, P. J., Kozłowska, K., Murphy, N. A., and Lowe, L. (2018). WordNet-feelings: A Linguistic Categorisation of Human Feelings. *arXiv preprint arXiv:1811.02435*.
- Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Wordnet As a Backbone of Domain and Application Conceptualizations in Systems with Multimodal Data

Jacek Marciniak

Department of Artificial Intelligence, Faculty of Mathematics and Computer Science,
Adam Mickiewicz University in Poznań
Uniwersytetu Poznańskiego 4 Street, 61-614 Poznań, Poland
jacekmar@amu.edu.pl

Abstract

Information systems gathering big amounts of resources growing with time containing distinct modalities (text, audio, video, images, GIS) and aggregating content in various ways (modular e-learning modules, Web systems presenting cultural artefacts) require tools supporting content description. The subject of the description may be the topic and the characteristics of the content expressed by sets of attributes. To describe such resources one can just use some of existing indexing languages like thesauri, classification systems, domain and upper ontologies, terminologies or dictionaries. When appropriate language does not exist, it is necessary to build a new system, which will have to serve both experts who describe resources and non-experts who search through them. The solution presented in this paper used to resource description, allows experts to freely select words and expressions, which are organized in hierarchies of various nature, including that of domain and application character. This is based on the wordnet structure, which introduces a clear order for each of these groups due to its lexical nature. The paper presents two systems where such approach was applied: the E-archaeology.org e-learning content repository in which domain knowledge was integrated to describe content topics and the Hatch system gathering multimodal information about the archaeological site targeted at a wide audience, where application conceptualization was applied to describe the content by a set of attributes.

Keywords: domain and application conceptualizations, wordnet based ontologies, multi-relational and multi-hierarchical indexing languages

1. Introduction

Before building an information system, it is necessary to make a decision regarding the way of organizing the information within so that the data supply process is simple and secure. In order to make the process run smoothly, it is necessary to select the solutions which will support describing objects of a similar kind in a consistent way. This is especially essential when data are input into the system by multiple users working in different time, because there is a risk of describing the same objects in many different ways. Moreover, during the data input various errors will appear, e.g. duplicated entries, incomplete or inconsistent data. In business systems this problem is noticed because of the big scale of this issue. Some researches show that nearly 40% of all company data is found to be inaccurate, or that for instance 92% of businesses admit their contact data is not accurate (Halo, 2020). This creates a need of data cleaning, which takes the form of standardization (replacing of different instances of the same value with one value) or deduplication (detection of duplicate values and their consolidation). These problems appear even when processing data as obvious as e.g. the recipient's address. Therefore, handling them will be a much greater challenge in the case of less obvious data like a type and a nature of pattern of a painting found at an archaeological site (i.e. zoomorphic, geometric, bucranium, wall painting). In such cases, data cleaning must be carried out by experts, who due to little amount of time and working in the project rigor will rarely be available when the data coherence processes will be necessary. Carrying the data cleaning processes out is always laborious and costly, so it is a wise idea to care about the data coherence when entering them into the system. In order to do this, existing dictionaries, terminologies, thesauri, classification systems or ontologies may be used. This solution may be useful when building systems which are at the advanced stage of the development cycle and store content of universal or well-developed area. Only in

such cases it can be assumed that there exists some available indexing language, which would support describing the content in a homogeneous way. Even then, we cannot be sure that all users will perform the process in the same way. Even when dictionary, thesauri or classification system or ontology are used, users can describe the resources in different ways (Hjørland, 2012). This means that they can describe the same object using different words, terms or classes from classification system. The situation is even more complicated when information system is at the initial stage of development and tools supporting resources description do not exist, or existing indexing languages do not comply with the needs due to e.g. cultural differences or domain conceptualization not concordant with the needs of experts responsible for describing the resources.

The paper presents the solution in which the data description is carried out using the indexing language being built during the process of the multimodal data input. The solution has been chosen due to the fact that prior to building two given information systems there was no dictionary, thesaurus, classification system or ontology which would be applicable in the resource description process. In the adopted approach, the experts who input multimodal content into the system, describe it at the same time using freely chosen words or expressions. They are organized into hierarchies and connected with the relations of different nature, including that of domain and application type. The solution is based on the wordnet structure and uses its hierarchy as the core organization of the developed indexing language. Two information systems in which this approach was used are: the E-archaeology.org e-learning content repository, where the content description is carried out using the lexical units taken from a wordnet and extended with a domain conceptualization, and the Hatch system storing multimodal data from archaeological site in Çatalhöyük. In the last case, the wordnet structure was supplemented with an application conceptualization.

2. Other solutions for data organisation

In information systems, data structures are in most cases the integral part of the system. An attribute-value approach is used, attributes are part of a system architecture and their organization is determined by programmers during system implementation. Data input by users are added to a database and they can be maintained in it. Other architectures, such as ontology-driven software architectures, allow modelling data structures outside the system (Pan et al., 2013). Such approaches allow to improve the exchange, maintenance and hierarchization of attributes and values assigned to them.

The common programmers' practice is validating data that are input to the system to avoid errors. In the simplest case, validation takes a form of checking the user input with the data type required for the attribute. The input may also be compared with internal dictionary entries. This solution is insufficient when the dictionary can be expanded by users inputting data, or when dictionaries from different systems need to be used. In such cases unexpected errors may appear, such as multiple entries describing the same concept or repeated values.

The solution to these problems is using existing dictionaries, terminologies, thesauri, classification systems, ontologies etc. when describing content (Crofts et al., 2010), (Gemet, 2020), (Getty AAT, 2020), (Geonames, 2020), (Iconclass, 2020), (Niles, Pease, 2001). The use of existing indexing language supports describing the content in a homogeneous way by multiple users. Yet, it forces indexers to refer to the existing conceptualization of the domain, which is why sometimes it may occur impossible to describe the content in a satisfactory way. Dissatisfaction may result from missing terms, hierarchization incompliant with expectations, granularity of concepts and habits of experts describing contents. If an existing controlled vocabulary or classification system is used, and there will be a need to change or add new descriptors to the existing language when indexing, the extension process may be excessively lengthy (Weda, 2016). At times, if the used language is developed by another team, the extension will not be possible at all. Among the problems with using controlled vocabulary to index the resources, there are also: difficulties in differentiating specific and general vocabulary, arbitrariness when defining synonymy and introducing abbreviations or acronyms to vocabulary, adding qualifiers when handling homographs, homonyms, different approach when introducing common and technical terms (Joudrey et al., 2018). Therefore, while dealing with content description, it is beneficial to use a language which allows for maintenance of different types of conceptualization, including the ones that can be extended during the description process and ones that cannot due to their controlled character.

Even if during the description process we use the existing indexing language such as dictionary, thesaurus or classification system, we must remember that access to indexed resources does not necessarily have to be easier (Maniez, 1997), (Hjørland, 2012). It means that during indexing resources stored in some repository, expert responsible for indexing will make arbitrary decisions regarding the use of a particular indexing language. Then, there is a possibility that when describing a concept, one will use more general terms despite the occurrence in a given language of specific terms that allow describing the

subject in more detail way. Therefore, there is a need for solution which allows to detect such practices easily and to make corrections without the risk of generating additional errors.

Among numerous approaches to resource indexing, there are some in which a wordnet was used. Princeton WordNet was used, e.g., in indexing the works of arts as complement to other three description systems: Getty AAT, Iconclass and ULAN (Holing et al. 2003). In the LT4EL project, a wordnet was used to index e-learning content stored in LMS Ilias system (Monachesi et al., 2008). The relations used in the solutions were wordnet hyperonymy relation, some relations from Dolce ontology and others from a domain ontology. Some works were also conducted towards mapping thesauri onto wordnets (Maziarz, Piasecki, 2018). plWordNet was also used to enrich a keywords database of the Polish Classification of Activities indexing language (Jastrzab, Kwiatkowski, 2019). In all of these solutions, existing wordnets and other indexing languages were used. Thus, indexers taking part in the content description process could only use descriptors available within those systems.

3. Wordnet enhanced by a domain conceptualization for indexing and searching repository of eLearning content

For the needs of describing the subject of e-learning content stored in the E-archeology.org repository, it was necessary to develop a solution which would allow organizing words and expressions used in the process of resource tagging in a way that supports indexing processes and searching through resources. The repository contains e-learning materials on the protection of archaeological heritage, the management and protection of cultural and natural heritage and introductory materials on archaeology for engineers and engineering for archaeologists (Marciniak, 2014). Currently, the repository contains more than 6,200 learning objects in 9 languages, which together create around 1,700 modules and units, and more than 30 training curricula (Marciniak, 2019a). The content includes text materials, graphics, films, quizzes and animations (Fig. 1).

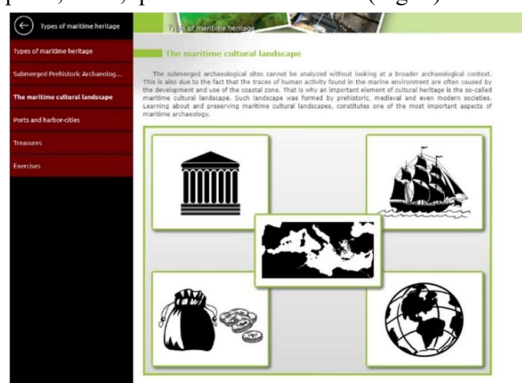


Figure 1: E-learning materials in the E-archeology.org content repository

The e-learning content stored in the repository is compositional and constructed in such a way that allows creating new training curricula from existing modules and units. Initially, the repository contained content regarding protection of archaeological heritage (Marciniak, 2014), and later the materials about management and protection of

cultural and natural heritage were added (Marciniak, 2019a).

Considering the large volume of the repository and the diversity of subjects, a proper description of the content is necessary in order to effectively search for modules and units, when new training curricula are compiled. The subject of contents was described by tagging (Smith, 2008). In this process words and expressions freely chosen by an indexer were stored in metadata assigned to e-learning components (keyword metadata from the IEEE LOM scheme). As in the tagging process, the indexers are not limited in terms of tags they use. It is necessary to organize them so that they can be re-used by other taggers. This will allow indexers to choose words and expressions of the appropriate level of detail when the system will propose more than one candidate to choose from.

In order to organize concepts by referring to the knowledge available only to experts, a conceptualization of archaeological and natural heritage domain was introduced into the created indexing language. The wordnet relations in it between words and expressions are intended to provide synonymy support and to allow a distinction of description detail (especially by use of hyperonymy relation) which will both be understandable for experts and non-experts.

3.1 The structure of expanded wordnet and its role in indexing and searching the repository

Words and expressions used during tagging e-learning content were then used to create the PMAH (Protection and Management of Archaeological Heritage) indexing language. At the initial stage of its development finished in 2015, it contained only words and expressions in English and it covered the domain of management and protection of archaeological heritage (Marciniak, 2016). Afterwards, along with providing the repository with a new content, the domain was expanded with management and protection of cultural and natural heritage. It was done by adding new words and expressions and a new domain hierarchy of concepts.

Among words and expressions used when tagging resources, we can distinguish common names (e.g. anthropology, aircraft, aerial archaeology), proper names (e.g. British Museum, Altamira), surnames (e.g. Eric Hobsbawm), geographical names (e.g. Gzira Stadium, France, Europe) and dates (e.g. 1956, 1940–1945).

For the purposes of facilitating the content tagging and searching process by recommendation of more tag candidates to system users (Fig. 2), the words and expressions were connected by the following relations:

- synset to consider the words or expressions as synonymous,
- wordnet relations between synsets (hyperonymy, holonymy, belongs to class),
- domain relations between synsets introduced by domain experts,
- generated relations between synsets determining similarity and relatedness of concepts,
- synsets assignment to domain categories determining the domain hierarchy.

The task of wordnet relations is to organize words and expressions in a way that is understandable to all repository users, not only to domain experts. Lexical relations are understandable for all users — both experts and non-experts. Connecting entries using wordnet relations is intended to help the users who do not know the specialized

terminology to select of the most appropriate tags when indexing and searching resources. When tagging resources by referring to relations such as hyperonymy / hyponymy (e.g. archaeology – aerial archaeology), holonymy / meronymy (e.g. cultural heritage – cultural heritage management), instance/class (e.g. Altamira - cave), experts can select the tags of an adequate level of detail, increasing the chance of using the tags previously used by other users. The synonymy relation is indicated as one of the basic types of relations used in indexing languages and appears, e.g., in the specification defining the thesauri form (Dextre Clarke, Lei Zeng, 2012). In contrast to controlled vocabularies such as thesauri, the use of synsets to describe synonymy makes indicating the descriptor, i.e. the preferred term impossible. In case of the approach in which indexing of resources takes the form of tagging, this is the expected characteristic of chosen solution. Currently, the words and expressions are grouped in c. 2000 synsets in the PMAH indexing language.

Domain relations between synsets were introduced by the domain experts in order to express the relations of an indefinite nature (e.g. archaeology – archaeological project, heritage – archaeological heritage protection). In the case of PMAH, the used relation was *link*. This relation refers to the fuzzynymy relation from wordnets (Vossen, 2002), (Maziarz et al., 2011) and associative relations from thesauri (Dextre Clarke, Lei Zeng, 2012). Introducing such relations is to allow indexers to access words and expressions connected within the domain. The set of relations between synsets was complemented with the relations defining similarity and relatedness of concepts, which are generated using heuristic rules (Marciniak, 2016). The rules refer to, *inter alia*, wordnet hierarchy (e.g. HasSameHypernym) and produce new relations between synsets to increase the number of tag candidates proposed by the system during tagging and searching through the repository (Fig. 2).

Search for:

The screenshot shows a search interface with three tabs: 'Simple', 'Tags', and 'Metadata'. The 'Simple' tab is active. Below the tabs, there are radio buttons for 'Used' and 'All', with 'All' selected. A search input field contains 'archaeological project' and a language dropdown is set to 'English'. Below the search bar, there is a 'Choose tag from ontology' dropdown menu showing 'PMAH Ontology'. To the right, there is a search button and a list of recommended tag candidates. The first candidate is 'archaeological project (1)' with a checkbox that is checked. Below this candidate, there are details: 'ID: 120', 'Categories: Process (1)', and 'Related: research project (1), work (1), archaeology (1), project (1)'.

Figure 2: Recommended tag candidates during tagging and searching through the repository

In addition to relations between synsets, all synsets were also mapped onto hierarchical structures created using so-called domain categories (DC). In the adopted approach, the domain categories perform the function of semantic labels used to represent the concepts derived from thesauri, classification systems or domain ontologies. They perform a function analogical to semantic domain from WordNet or domain labels from EuroWordnet allowing a proper organization (categorization) of synsets and being used to group synsets of one semantic field (Felbaum, 1998),

(Vossen, 2002). The hierarchical structure of domain categories is created using the generic or mereological relations. The conceptualization obtained by means of domain categories hierarchy and synsets mapped onto them, is of the multi-faceted character. Initially, 12 the most general domain categories were placed at the top of the domain categories hierarchy. When words and expressions from new subject domains were used as tags during uploading the contents of different subject into the repository, the number of domain categories at the top of the hierarchy increased to 26. Among them, there are categories like Archaeological heritage, Archaeological process, Chronology, Archaeology, Landscape, Nature, Policy, etc. Now, the number of all domain categories is 238. The hierarchy of domain categories with assigned synsets is presented to users searching and tagging the repository as a hierarchical index (Fig. 3).



Figure 3: Index of tags in the domain categories hierarchy

The structure of the wordnet hierarchy extended with domain relations and domain categories hierarchy is presented in Fig. 4. The indexing language created in such a way can be considered as an ontology understood as an arrangement of objects appearing in a given domain and the knowledge about them shared by specialists or as a specification of conceptualization (Joudrey et al., 2018), (Gruber, 1993). The wordnet based ontology thus understood, following Uschold's and Gruninger's (1996) formalization, will be considered as a semiformal ontology.

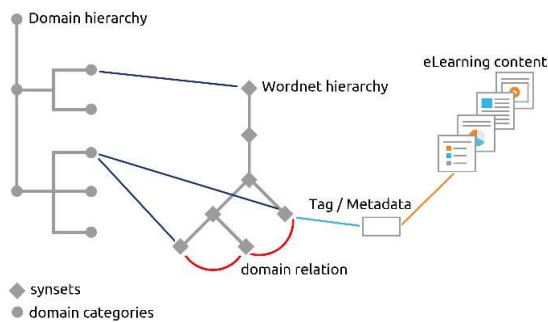


Figure 4: The structure of the wordnet hierarchy extended with domain relations and domain categories hierarchy

3.2 Building extended wordnet while tagging eLearning resources

The PMAH ontology was developed along with the expansion of content in the E-archaeology.org repository. At the early stage of the development, the initial set of words and expressions (c. 1,000) used by experts in the process of content tagging was then expanded by additional 400 entries (synonyms, more general terms and terms connected with associative relations) (Marciniak, 2016). The hierarchical structure for these entries was developed on the basis of the existing wordnet structure (i.e. Princeton WordNet) considered as a referential wordnet according to the algorithm of wordnet based ontology creation (Marciniak, 2016). In the case of the PMAH ontology, the algorithm aimed to integrate all words and expressions used by taggers into ontology. It expanded the ontology only in those fragments in which a new synset was included. It did not aimed to incorporate all synsets from the referential wordnet, only hyperonyms of the new introduced synset were added. According to the algorithm, domain relations (i.e. associative relations) between synsets were added by domain experts. They also created the hierarchy of domain categories and mapped synsets onto them.

At the second stage of development, when the repository was expanded with the content from management and protection of cultural and natural heritage domain, additional 600 words and expressions were added into the PMAH ontology. At this stage, the process of adding all new words or expressions to the ontology took place directly during tagging e-learning materials. Because the ontology was already built and contained a substantial set of entries, the system suggested to an indexer words or expressions used earlier in the repository as tags by other indexers (Fig. 5).



Figure 5: Tagging resources in the repository

If the indexer (an expert), did not found a candidate to be used as a tag among words and expressions from the ontology, he or she always could add a new tag. Such a tag was assigned to e-learning content metadata and added at the same time to the PMAH ontology. This process was performed in two steps:

- the expert's task was to assign the word or expressions to one or multiple domain categories, add synonyms or associative relations with other synsets from the ontology,
- a lexicographer added afterwards the unit to the wordnet hierarchy.

The first actions were undertaken in the content repository at the time of content tagging with the use of one combine form (Fig. 6). The expert could choose domain categories onto which the introduced tag had to be assigned to, as well as word or expression from the ontology to be connected with the associative relations.

The latter actions were performed outside the e-learning content repository in a dedicated tool - the Ontology Repository Tool (Marciniak, 2019b). Using the external tool allowed the introduction of necessary modifications and extensions into the ontology, such as typo corrections, removal of duplicates or hierarchy adjustments. It allowed to carry out ontology maintenance processes by knowledge engineers (domain experts) who did not needed to be supported by programming teams.

Figure 6: Adding an expression to the domain structure of the PMAH ontology and linking it to a synset

4. Wordnet enhanced by an application conceptualization for describing artefacts of heterogenous character

The second application of the solution based on wordnet structure enhanced by the expert knowledge, was the use of the extended PMAH ontology in the Hatch system. The Hatch (House at Çatalhöyük) is an advanced Web system designed to create and maintain a digital collection (Marciniak et. al, 2020). The Hatch is aimed at presenting a wide range of multimodal data about the Neolithic settlement at Çatalhöyük in a multiscalar and interactive form. It combines information of different character (types of artefacts, their attributes, relations among them) with different form of their presentation (text, photographs, graphics, maps, GIS localizations and multiscalar chronology of artefacts). It is designed to meet the needs and expectations of both professionals and general public interested in the human past (Fig. 7).

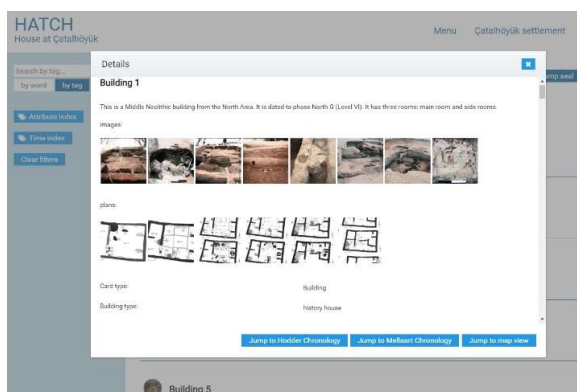


Figure 7: A card of an artefact in the Hatch system

The system was constructed when the excavation works at the site were very advanced. At the time of the system construction, the expert team already had a large amount of various data about the site, such as text descriptions, photographic material, maps, GIS database and artefacts chronology. Yet, the data were not organized in terms of their presentation in a system for users without specialist knowledge about Çatalhöyük site. Due to the character of the archaeological site, there was no indexing language which could be used to describe the resources stored in the Hatch system. Therefore, a solution was adopted in which the PMAH ontology was extended with entries related to the Neolithic site, with consideration to the character of Çatalhöyük. Furthermore, the Hatch system is to be supplemented with e-learning courses, which will supply the E-archaeology.org repository and will have to be tagged in a similar way to other resources stored there.

In contrast to tagging e-learning content in the repository, where all words and expressions chosen as tags by indexers are assigned to one metadata, entries from the PMAH ontology are assigned as a values to multiple attributes describing artefacts in the Hatch system. The number of attribute organization schemes equals the number of object types stored in the Hatch. Their arrangement results from the need to present the data in the system and that is why it has simply applicational character.

4.1 The structure of extended wordnet and its role in describing artefacts of different type

Words and expressions which extended the PMAH ontology were obtained during data input into the Hatch system. The artefacts are organized in the system, in so-called cards, where attribute-value structures serve to describe artefacts' characteristics. The number of attribute-value pairs is different for each object type and the corresponding card. For instance, the attributes for an Imagery card are *Imagery type* and *Motifs*, respectively taking exemplary values of *wall painting* and *zoomorphic*. In Animal bones card for *Animal bones types* attribute, the exemplary values are *astragali*, *crane ulna* or *scapula*. Attribute-value structures were constructed using a new domain category type and the synsets assigned to them. The new domain category (DC-HATC) is different than the one used in the previously presented solution used for tagging e-learning content, because the character of a new hierarchical arrangement of concepts in the PMAH ontology built to accomplish the Hatch system needs, is also different. The approach in which attribute-value structures are built with domain categories embedded in the ontology, make possible the storage and maintenance of the data outside the Hatch system. It implements the postulate of getting the information structures out from the information system, which streamlines the process of correcting words and expressions used for resource indexing.

The fact that information structures are hosted outside the information system facilitates the use of the same word or expression as values assigned to several attributes. This creates a possibility to reuse the word or expression which was used before as the value in a different attribute. For example, the *zoomorphic* value was used as a value of two attributes describing the motif type: in Stamp seal card and Imagery card. Thanks to this, the user searching through the system will receive the cards of two different types when typing the *zoomorphic* value as the query to the system.

Due to the character of the archaeological artefacts for which words and expressions were used as attributes' values, the words and expressions used in the Hatch system can be divided into:

- units equivalents of which can be found in the largest reference wordnets, e.g. plWordnet (plWordnet, 2020) or Princeton WordNet (WordNet Search, 2020), e.g. bucranium, flint, geometric, kerb, relief,
 - units for which equivalents could not be found in any referential wordnet, including units which could be and those which couldn't be added there for different reasons e.g. astragali, animal bone, crane ulna, abandonment deposit, zoomorphic,
 - units with a strong terminological character, e.g. barley seeds, feasting deposit, post retrieval pit, multi-roomed construction,
 - expression referring to the time, e.g. "3–12 years – child", "20+ adult",
 - chronology in qualitative units (TP M, Level II, North I).
- For the purposes of the Hatch system, the words and expressions were connected by the following relations:
- synsets connecting words and expressions considered to be synonymous,
 - wordnet relations between synsets (hyperonymy, holonymy, belongs to class),
 - generated relations between synsets determining similarity of concepts,
 - synsets assigned to domain categories,
 - special relations for handling object dating,

In the Hatch system, synonymy relation was used to keep the information about the singular or plural form of words or expressions used as descriptors. There is no general rule regarding the use of singular or plural in descriptors. It depends on the specific language and the regulations adopted by the individual community or country (Joudrey et al., 2018). In the Hatch system, in a situation where singular and plural was used as values of the same attribute, this was not considered as an error and was not corrected. Instead, both forms were related in one synset. The solution is not canonical and was adopted because of the practical matters. In the process of synsets creation, an interesting problem of ambiguity arose. For instance, in the case of a word building it was necessary to make a decision whether it fulfils the definition from the referential wordnet, or it is necessary to introduce a new meaning and create a new synset due to the character of the Neolithic buildings located at the Çatalhöyük site. The first solution was chosen, despite it may be debatable in the case of domain and applicational uses of the PMAH ontology.

Wordnet relations were used to relate those words and expressions (synsets) which were found in the referential wordnet, as well as for those which could not be found. This approach was adapted due to the need of the rules generating relations between synsets determining concepts similarity, which use lexical relations, especially hyperonymy relation. As in the case of the e-learning content repository, the generated relations determining similarity and relatedness between synsets are intended to be used in recommendation of best tag candidates to the Hatch non-expert users searching the system.

Similarly to the e-learning content repository, all synsets were mapped onto hierarchical structures built using domain categories. Due to a different character of this hierarchy, other type of category was used (DC-HATC). This hierarchical arrangement of words and expressions is

useful only in the case of the Hatch system because of its strongly applicational character. At the top of the domain categories hierarchy, there are three categories which arrange words and expressions considering their role in the Hatch system: Attributes, Auxiliary attributes and Time Index. Other domain categories being attributes of cards (Animal bones, Figurine, Imagery, Pottery, etc.) are subcategories of the Attributes category. In general, there are 57 domain categories arranging words and expressions taking the Hatch needs into account. Domain categories hierarchy with assigned synsets is presented to the users searching through the Hatch system as two separate hierarchical indexes: Attribute index (Fig. 8) and Time index (Fig. 9).

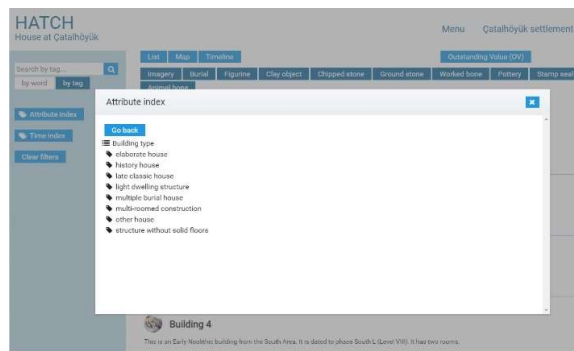


Figure 8: Attribute index in the Hatch system

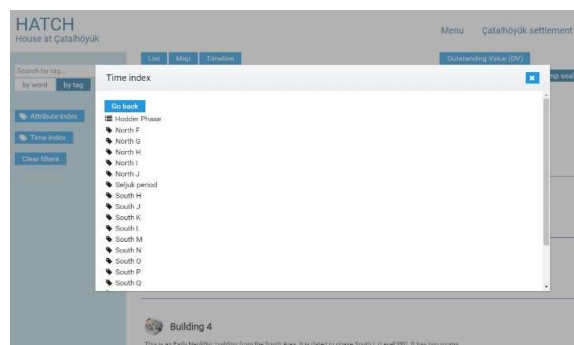


Figure 9: Time index in the Hatch system

Time index shows a special use of relations from the PMAH ontology for chronology arrangement of artefacts at the Çatalhöyük site. Due to the character of the site, chronological order is arranged with qualitative values. Absolute dating using C14 method is available only for selected objects. Therefore, when presenting the chronology of the objects in the Hatch system, three different systems developed for the needs of Çatalhöyük site were used: Mellaart Phase, Hodder Phase and TP Phase. Each system consists of a set of highly terminological values, e.g. North F, Level III, TP M. As the timeline with artefacts from the site is one of the ways of presenting the objects in the Hatch system, it was necessary to assign qualitative values used in the chronology system to particular dates, so that the date can be interpreted in a programming component used to create the timeline. As in the case of other values assigned to attributes, terms from a chronology system (e.g. North F) are also assigned to domain categories from the PMAH ontology. Those entries

were connected with dates (e.g. 6300 BC) specifying the approximate and conventional (from the point of view of the archaeological research methodology) time of a given period. The relations used have associative and applicational character, i.e. they are not useful outside the Hatch system.

The structure of the wordnet hierarchy extended with domain and applicational hierarchies is presented in Fig. 10.

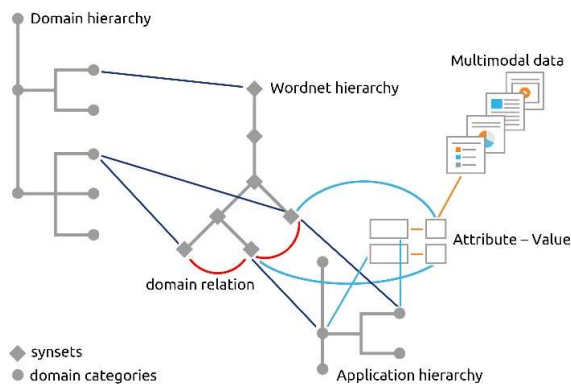


Figure 10: The structure of the wordnet hierarchy extended with domain and application hierarchies

4.2 Building extended wordnet while describing multimodal data

The PMAH ontology was extended for the needs of organizing the information in the Hatch system when the system was fulfilled with the multimodal data such as photos, maps, GIS data, text descriptions and bibliographic references. They were grouped into cards corresponding to different artefacts types. In total, 725 cards, 1107 photos, 194 maps and 71 000 GIS objects were input into the system. The process of supplying the system with the data was carried out by a few domain experts for about a year. For the domain experts (archaeologists), it was mainly the task of ordering information about artefacts from the site, choosing appropriate photographic materials, locating the object on the GIS map and determining the chronology of artefacts. Assigning words or expressions as values of attributes was performed simultaneously to other actions and was not prominent. As there was a risk of errors appearing in the process, values were assigned to attributes in one form directly in the Hatch system. Its goal was to minimize the number of errors appearing when several experts were extending the PMAH ontology at the same time. The goal was achieved when assigning values to attributes due to (Fig. 11):

- suggesting by the system words or expressions which were used before by other indexers as a value in an attribute,
- suggesting by the system words or expressions which were not used before as a value in the attribute, but which were present in the ontology due to the fact that they were either assigned as a value to another attribute before, or were just present in the ontology, but not yet used in the Hatch system,
- entering new words or expressions and assigning them as a value to a particular attribute.

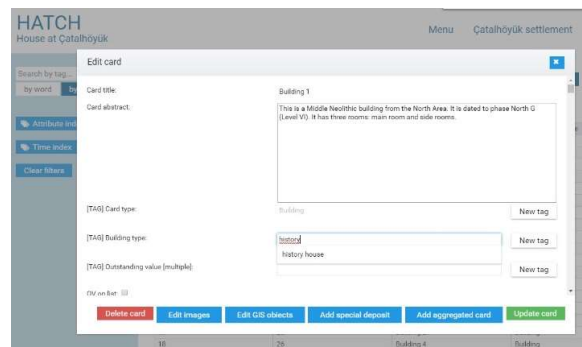


Figure 11: Adding a word to the domain structure of the PMAH ontology during entering data in the Hatch system

Words and expressions entered during artefacts description by domain experts, as well as the PMAH ontology, were placed in the external tool, which was used for maintenance tasks. The maintenance of the data was periodically handled by one domain expert, who controlled the entered words and expressions and introduced corrections such as deleting the entries with errors (e.g. typographic errors), deleting the values inconsistent with the description criteria adapted by the team, replacing too general or too detailed values and the ones of an inappropriate granularity. Other deleted elements included incorrect values resulting from the software engineering errors and internet connection errors.

5. Conclusion

The solution presented in this paper shows that in the process of indexing resources of different character and highly specialized subject, it is necessary to use indexing languages which allow to extend them according to the needs with maintaining the clear organization of terms at the same time. Application of wordnet based ontology using the wordnet structure as a backbone of the whole system, allows to use arrangement resulting from the wordnet and refers to conceptualization available for both experts and non-experts. Due to such structure, a non-expert will be able to switch between specialized terminology and words and expressions known from common language, thanks to the tag candidates recommendation facility available in the presented systems. This will allow non-experts to formulate more appropriate queries when searching through the repository. Experts will be able to choose the most appropriate level of detail when indexing a resource. Incorporation of domain and applicational conceptualizations to the system allows distinguishing different arrangement of terms meeting different needs in one indexing language. Domain ordering allows experts to arrange entries according to their specific needs and knowledge. Applicational ordering improves the process of resource description, as it allows using words and expressions already used before for indexing resources by other experts.

Due to the separation of knowledge structures outside the system in which they are used, it is possible to carry out ontology maintenance processes by knowledge engineers who do not need to be supported by programming teams. This makes the ontology maintenance process more clear and keeps the indexing consistent, when the action is performed by multiple users. This creates a possibility to

introduce changes and extensions to the indexing language without changing the IT structure of the system in which this indexing language is used.

6. Bibliographical References

- Crofts, N., Doerr, M., Gill, T., Stead, S., Stiff, M. (ed.) (2010). Definition of the CIDOC Conceptual Reference Model, ICOM/CIDOC.
- Dextre Clarke, S.G., Lei Zeng, M. (2012). From ISO 2788 to ISO 25964. The Evolution of Thesaurus Standards towards Interoperability and Data Modeling, *ISO Information Standards Quarterly*, Winter 2012, vol. 24.
- Fellbaum, Ch. (ed.) (1998). *WordNet: An Electronic Lexical Database*, MIT Press.
- Getty AAT. (2014). About the AAT, . Access data: February 2020.
- Gemet. (2012). Gemet: General Multilingual Environmental Thesaurus, <http://www.eionet.europa.eu/gemet/en/about/>, Access Date: February 2020.
- Geonames. (2020): The GeoNames geographical database, <http://www.geonames.org>, Access date: February 2020.
- Halo. (2020). Data Quality in BI the Costs and Benefits, <https://halobi.com/blog/infographic-data-quality-in-bi-the-costs-and-benefits/>, Access data: February 2020.
- Iconclass RKD (2020): IconClass, www.iconclass.nl, Access Date: February 2020.
- Hjørland, B. (2012). Is classification necessary after Google?, *Journal of Documentation*, vol. 68, iss. 3, pp. 299-317.
- Hollink, L., Schreiber, G., Wielemaker, J., Wielinga, B. (2003). Semantic annotation of image collections. S. Handschuh, M. Koivunen, R. Dieng, S. Staab (eds.), *Proceedings of the KCAP'03 Workshop on Knowledge Capture and Semantic Annotation*, Florida, October 2003, s. 41-48.
- Jastrząb, T., Kwiatkowski, G. (2019). Enriching a Keywords Database Using Wordnets – a Case Study. *Proceedings of the 10th Global Wordnet Conference*, Wrocław, July 23-27 2019, pp. 329-335.
- Joudrey, D.N, Taylor, A.G., Wisser, K.M. (2018): *The Organization of Information*, 4th ed. Libraries Unlimited.
- Maniez, J. (1997). Database merging and the compatibility of indexing languages, *Knowledge Organization*, vol. 24, no. 4, s. 213-224.
- Marciniak, A., Marciniak, J., Filipowicz, P., Harabasz, K., Hordecki, J. (2020). Engaging with the Çatalhöyük database. House at Çatalhöyük (HATCH) and other applications. *Near Eastern Archaeology*, 83:2.
- Marciniak, J. (2014). Building E-learning Content Repositories to Support Content Reusability, In *International Journal of Emerging Technologies in Learning (iJET)*, Volume 9, Issue 3 (2014), pp. 45-52.
- Marciniak J. (2016). Building wordnet based ontologies with expert knowledge. Zygmunt Vetulani, Hans Uszkoreit, Marek Kubis (ed.) *Human Language Technology. Challenges for Computer Science and Linguistics Papers, Lecture Notes in Computer Science*, Vol. 9561, pp. 243-254, Springer International Publishing.
- Marciniak J. (2019a). Methods and Tools for Centers of Integrated Teaching Excellence Providing Training in Complementary Fields. *Proceedings of 11th International Conference on Computer Supported Education (CSEDU 2019) - Volume 2*, pp. 527-534.
- Marciniak J. (2019b). Ontology Repository Tool for effective development and deployment of wordnet based ontologies. *Proceedings of 9th Language and Technology Conference (LTC 2019), Human Language Technologies as a Challenge for Computer Science and Linguistics - 2019*, pp. 25-26.
- Maziarz, M., Piasecki, M., Szpakowicz, S., Rabiega-Wiśniewska, J. (2011). Semantic relations among nouns in Polish Wordnet grounded in lexicographic and semantic tradition, *Cognitive Studies*, vol. 11.
- Maziarz, M., Piasecki, M. (2018). Towards Mapping Thesauri onto plWordNet. *Proceedings of the 9th Global Wordnet Conference*, Singapore, 8-12 January 2018, pp. 45-53.
- Monachesi, P., Simov, K., Mossel, E., Osenova, P., Lemnitzer, L. (2008). What ontologies can do for eLearning. *Proceedings of IMCL 2008*. 16-18 April 2008.
- Niles, I., Pease, A. (2001). Towards a Standard Upper Ontology, *Proceedings of the International Conference on Formal Ontology in Information Systems (FOIS'01) – Volume 2001, October 2001, ACM*, pp. 2-9.
- Pan, J., Staab, S., Aßmann, U., Ebert, J., Zhao, Y. (eds.) (2013). *Ontology-Driven Software Development*. Springer, 2013
- plWordnet (2020): plWordnet, Słowsociec, a large network <http://plwordnet.pwr.wroc.pl/wordnet/>, Access Date: February 2020.
- Uschold, M., Gruninger, M. (1996). Ontologies: principles, methods, applications, *Knowledge Engineering Review*, vol. 11(2), s. 93-136.
- Weda, R. (2016). Update on the Dutch AAT work. https://www.getty.edu/research/tools/vocabularies/weda_zelfde_aat_dutch_2016.pdf, Access : February 2020.
- WordNet Search (2020). Wordnet Search 3.1. <http://wordnetweb.princeton.edu/perl/webwn>. Access : February 2020.
- Vossen, P. (ed.) (2002). *Euro WordNet General Document. Version 3*, University of Amsterdam.

Inclusion of Lithological terms (rocks and minerals) in The Open Wordnet for English

Alexandre Tessarollo, Alexandre Rademaker

Petrobras, IBM Research and FGV/EMAp
alexandretessarollo@gmail.com, alexrad@br.ibm.com

Abstract

We extend the Open WordNet for English (OWN-EN) with rock-related and other lithological terms using the authoritative source of GBA's Thesaurus. Our aim is to improve WordNet to better function within Oil & Gas domain, particularly geoscience texts. We use a three step approach: a proof of concept-level extension of WordNet, a major extension on which we evaluate the impact with positive results and a full extension encompassing all GBA's lithological terms. We also build a mapping to GBA which also links to several other resources: WikiData, British Geological Survey, Inspire, GeoSciML and DBpedia.

Keywords: wordnet, rocks, lithology, domain extension, geology, NLP

1. Introduction

Oil & Gas Exploration and Production companies annually invest billions of dollars gathering documents such as reports, scientific articles, business intelligence articles and so on. These documents are the main base for major decisions such as whether to drill exploratory wells, bid or buy, production schedules and risk assessments (Rademaker, 2018). However, most of the processing of this fundamental data is still done by human professionals actually reading it rather than by a computational system. Considering that this unstructured data is growing exponentially, management of such data and finding relevant content quickly has become one of companies and professionals most critical challenges (Antoniak et al., 2016; Schoen et al., 2018). Even though Natural Language Processing (NLP) has significantly advanced over the past years, the specific domain of Oil & Gas has its own challenges, some of them presented in (Rademaker, 2018).

Assessing geosciences papers one can notice that among the most common properties raised are usually geographic location (Palkowsky, 2005), geological time and lithological information. In a previous work (Rademaker et al., 2019) we addressed some of the issues regarding geological time. In this work we approach the lithological information aspect.

Section 2. gives a brief description of similar projects. Section 3. present our authoritative source for terms and definitions. Section 4. shows our platform of choice for extending the WordNet. In section 5. we present and discuss the proposed changes. In section 6. we raise some relevant and recurrent issues we faced and the reasoning supporting our decisions. Section 7. presents some comparative statistics over a given corpus processed both with the original WordNet and our extended version. Section 8. sums up the results and points to future works.

2. Related works

Princeton WordNet (PWN) (Fellbaum, 1998a) does not cover many terms and concepts specific to certain domains as pointed out by (Buitelaar and Sacaleanu, 2002), hence the need to expand PWN for each domain in order to tap into its potential as a NLP resource (Amaro and Mendes,

2012). WordNet extensions for specific domains are relatively common.

Medical WordNet (MWN) (Smith and Fellbaum, 2004) reviews PWN medical terms through a corpus which includes a validated corpus of sentences involving specific medically relevant vocabulary. The corpus is composed by the definitions of medical terms already existing in WordNet, sentences generated via the semantic relations in PWN and sentences derived from online medical information services targeted to consumers. BioWN (Poprat et al., 2008) was another attempt to extend WN to the biomedical domain from the Open Biomedical Ontologies (OBO). OBO would provide terms, definitions and relations to be included in WN. According to the authors, the attempt failed due to issues on several softwares and resources that eventually prevented the success of the initiative. (Buitelaar and Sacaleanu, 2002) leans on German's compositional aspect to extend GermaNET with medical terms. The relevance of the candidate terms is then measured in a given domain corpora. Roughly the definitions arise from the compositional rule used to build the term in the first place.

In the legal domain, JurWN (Sagri et al., 2004) builds upon the Italian ItalWordNet (IWN) database, aiming to extend it to the legal domain. IWN (Roventini et al., 2003) is the Italian component of the EuroWordNet (Vossen, 2002). Words were selected from frequent terms used in queries of the major legal information retrieval systems, while definitions were taken from handbooks, dictionaries, legal encyclopedias and other main technical concepts. The LOIS (Lexical Ontologies for legal Information Sharing) project (Peters et al., 2006) encompass legal WordNets for six different languages (Italian, Dutch, Portuguese, German, Czech, English) based on the EuroWordNet framework. It used a subset of JurWN as a seed and added new terms on the basis of authoritative resources, national and EU legislative text and legal text.

GeoNames WordNet (GNWN) (Bond and Bond, 2019) links the GeoNames¹ geographical database to wordnets in different languages. GeoNames provides both the terms and definitions to be included in GNWN as an instance of a given synset (e.g.: Paris as an instance of city).

Noticeable from all these initiatives is the approach consid-

¹<https://www.GeoNames.org/>

ered to extend a wordnet to a given domain. Some refer to a corpus (custom built or pre-existing material) to gather a list of words to include in the wordnet, and then to an authoritative material such as dictionaries and encyclopedias for the definitions. Others refer to authoritative material that have both terms and definitions, such as ontologies.

3. INSPIRE and GBA's Thesaurus

The Infrastructure for Spatial Information in the European Community (INSPIRE) (Parliament and of the Council, 2007) was created to build upon existing resources (infrastructure and data) of the Member States. The original focus is to support EU policies and activities which may have an impact on the environment. Particularly within the scope of this work, Inspire offers an organized codelist for lithology². This resource is actually maintained by the Geological Survey of Austria (Geologische Bundesanstalt) within its "GBA Thesaurus" (GBA). Regarding lithology, GBA presents a richer material than Inspire, all accessible online³ and available for download⁴.

GBA is an ontology based on the Simple Knowledge Organization System (SKOS) vocabulary (Isaac and Summers, 2009). Each term has a Universal Resource Identifier (URI) and is related to other terms via SKOS object properties. Within the scope of our work, we have *broader* and its counterpart *narrower*. Therefore, "mammal has *broader* animal" and "animal has *narrower* mammal". GBA follows SKOS convention to only assert direct hierarchical links. The name of the term is given by *prefLabel* data property, while the definition is given by *definition* data property. String values are given in English as well as in German. GBA uses a few other SKOS properties like *related match*, *close match*, *hidden label* and others. Particularly *exact match* is used to map GBA to other resources, INSPIRE included. The downloadable material for GBA is a Resource Description Framework⁵ (RDF) file, which means it is organized in triples consisting of subject, predicate and object.

At its description, GBA states that Lithology comprises loose- and bed-rock, classified according to their modal composition and grain size, respectively. Magmatic-, polygenetic-, metamorphic- and fault-rocks are classified based on International Union of Geological Sciences (IUGS) recommendations⁶. Sedimentary rocks classifications refer to international standards. Considering GBA alignment with IUGS recommendations and its mapping to WikiData⁷, British Geological Survey (BGS)⁸, Inspire⁹,

²<http://inspire.ec.europa.eu/codelist/LithologyValue>

³<https://thesaurus.geolba.ac.at>

⁴<https://github.com/schmar00/gba-thesaurus/tree/master/rdf>

⁵<https://www.w3.org/TR/rdf-concepts/>

⁶<https://www.iugs.org/history>

⁷https://www.wikidata.org/wiki/Wikidata:Main_Page

⁸<http://data.bgs.ac.uk>

⁹<http://inspire.ec.europa.eu/codelist/LithologyValue>

GeoSciML¹⁰ and DBpedia¹¹, i.e. several governmental, multinational and community consensual based open-source initiatives, we assumed GBA's thesaurus for lithology as an authoritative figure. Therefore, it is not scope of this work to question the correctness of GBA's material, but to map it into the WordNet.

4. Princeton WordNet and the Open Wordnet for English

Princeton WordNet (PWN)¹² (Fellbaum, 1998b; Miller et al., 1990) is a large lexical database of English and one of the most widely-used language resources in natural language processing. It works well as a dictionary and a thesaurus for uses of English, as found, for instance, in newspapers and general knowledge texts, such as Wikipedia. Unfortunately, its development came to a halt over a decade ago.

In (Muniz et al., 2018) some of the authors present previous initiative to expand PWN with geological terms. This work started as fork of PWN release 3.0. Initially, PWN was converted to a human-readable text format and later an Emacs¹³ mode and a validation tool were developed. It is called Open Wordnet for English (OWN-EN) and maintained at <http://github.com/own-en/>. The focus is on the expansions of PWN to specific domains (mainly geology and its intersection with Oil & Gas exploration) but also on the fixing of well-known bugs founded in PWN over the years. In this repository one can find the products of this paper, i.e., the extended WN as well as the mapping between it and GBA.

In the future, we aim to consider the merge of our OWN-EN with the Open English WordNet (McCrae et al., 2019). This is another fork of PWN being developed under an open source methodology. Its 2019 release fixed over 3,500 errors in PWN. The authors are committed to release new versions at least every year. One can contribute to the project and/or use its products at <https://en-word.net>.

5. Extending OWN-EN from GBA's Thesaurus

WordNet's cornerstone is its several types of conceptual relations. Of our interest, we have the *hyponym of* (counterpart *hypernym of*), which indicates a subtype relation. The *part holonym of* (counterpart *part meronym of*) indicates a component relation. Similarly, *substance holonym of* (counterpart *substance meronym of*) indicates a component relation for substances. The *Domain of synset - topic* (counterpart *domain of synset - member*) indicates the topic a given concept (synset), as in "geology is *domain of synset - topic* of rock".

From the GBA thesaurus, we consider the labels and definitions of the concepts and the concepts relations. But GBA's definitions were not taken literally since they were

¹⁰<http://resource.geosci.ml.org/classifier/cgi/lithology>

¹¹<https://wiki.dbpedia.org>

¹²<https://wordnet.princeton.edu>

¹³<https://www.gnu.org/software/emacs/>

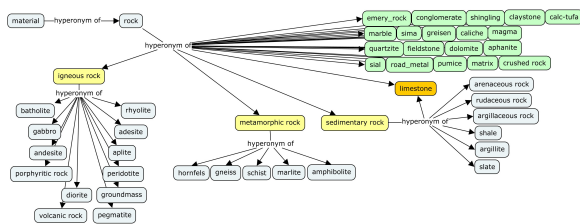


Figure 1: Rock in WordNet

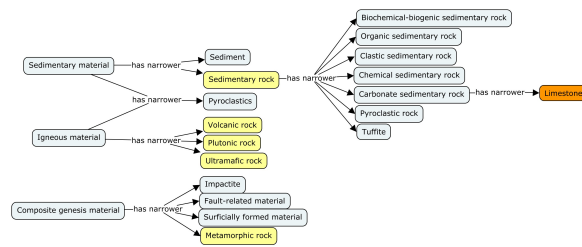


Figure 2: Rock in GBA

not written as dictionary definitions. For instance, they include many in-depth descriptions of the concepts and references to scientific literature. Our goal was to provide for the new synsets, as much as possible, Aristotelian definitions following general lexicography methodology. Besides all information from GBA incorporated into our OWN-EN, we also provide a mapping from GBA concepts URIs to the OWN-EN sense keys. This will also facilitate future revisions of our resource once new releases of GBA are made available. Because GBA is already mapped to multiple other resources (WikiData, BGS, Inspire, GeoSciML and DBpedia), our mapping encompasses these resources as well.

In WN, the word *rock* has many senses, and the one that resembles the geological meaning is 14696793-n (*rock* : material consisting of the aggregate of minerals like those making up the Earth's crust). The reader should consider this sense wherever *rock* is mentioned henceforth. Figure 1 shows how *rock* is represented in WN, while figure 2 shows a few of the uppermost lithologies in GBA. A first look at both shows that WN has at least some hierarchical issues: there are nineteen synsets (in green) that are *hyponym of rock* instead of one of the three main WN's classes of rock: igneous, metamorphic and sedimentary (all in yellow). Finally, there is *limestone* (in orange): *hyponym of both rock and sedimentary rock*. Considering *sedimentary rock* is *hyponym of rock*, the *limestone to rock hyponym of* is at least redundant.

In yellow in figure 2 we can see that *sedimentary rock* and *metamorphic rock* are represented in both WN and GBA. WordNet's *igneous rock* has three counterparts in GBA: *volcanic rock*, *plutonic rock* and *ultramorphic rock*. Finally, *limestone* in GBA is *hyponym of carbonate sedimentary rock* which in turn is *hyponym of sedimentary rock*. Notice that GBA does not have a term for 'rock' pure and simple. Instead its top concepts are three types of material and from those arise different rocks and other materials. 'Rock' however is used to define other ones (see *sedimentary rock* below). Due to this and to the fact that *rock* is a relevant term in everyday language, we chose to keep this WN synset, add the three top concepts of GBA and allocate GBA's specific terms downwards from these four synsets.

To expand and adapt WN onto lithology domain we used GBA's terms and properties starting from the different types of rocks and lithologies. The obvious choice for mapping SKOS relationships to WN relationships is as first discussed in (van Assem et al., 2006). In our case, where in GBA A has broader B, in WN we defined A as *hyponym of B*; likewise, where in GBA B has narrower A, in WN

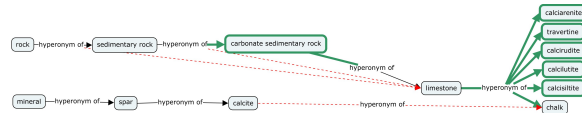


Figure 3: Limestone relations in WN: red ones to be removed, green ones to be included

we defined B as *hypernym of A*. For the sake of simplicity, we'll use WN's relations names henceforth. We also opted for lower case terms when changing or adding synset in WN.

GBA does not have explicit relations between rocks and the minerals that compose it, but we inferred the rock compositions in WN relations *substance holonym of* and *substance meronym of* from GBA's definitions. We also used WN's *domain of synset - TOPIC* and *member of this domain - TOPIC*, as explained later on.

As a proof of concept of our approach, we worked with *limestone* and initially analyzed only definitions and *hypernym of* and *hyponym of* relations. Afterwards we worked on the *substance holonym of*, *substance meronym of*, *domain of synset - TOPIC* and *member of this domain - TOPIC* relations. While the first step enriches WN with lithological terms, the second step ventures into the mineral domain, expanding WN even further. Once we set this work routine, we expanded the task to include all *carbonate sedimentary rock* and *clastic sedimentary rock*, the main types or reservoir rocks for Oil & Gas, ergo the most relevant for this industry. Finally, we included all of GBA lithology ontology into WN.

In WN *limestone* has the aforementioned redundant relations between *rock* and *limestone*. These and other deleted relations are highlighted in red in figure 3. In green the inclusion of 6 new terms and their 18 new relations with other terms. Note that due to the inclusion of *carbonate sedimentary rock* between *sedimentary rock* and *limestone* the *hypernym of* and *hyponym of* relations between *limestone* and *sedimentary rock* are no longer necessary.

For the six new terms added to WN we used the GBA definitions with minor adjustments in order to get closer to Aristotelian definitions and general lexicography methodology. For the ones that already existed in WN, a careful analysis was necessary and carried out top to bottom.

In GBA the concept *sedimentary rock* is defined as a *rock formed from post depositional consolidation of sediments (by processes of compaction, cementation, crystallization, or biogenic binding)* and it is a *hyponym of sedimentary*

material¹⁴. Analyzing both definitions and comparing with WN's definition for *sedimentary rock*¹⁵ we conclude that, as explained in Section 6., WN's current definition for *sedimentary rock* is technically poor and should be replaced.

The other words WN already had were *limestone* and *chalk*. *Chalk* was classified as a mineral in WN, but GBA states that *chalk* is a rock and that rocks are composed of minerals. WN had 14806598-n (*chalk* : a soft whitish calcite), while GBA defines it as *a light-coloured (white-gray) marine limestone composed almost entirely of fine crystalline calcite. These porous limestones consist of foraminifera and calcareous algae, and usually contain chert nodules.* On this term we discarded WN's current definition and replaced it with GBA's.

As for *limestone* WN has 14936226-n (*limestone* : a sedimentary rock consisting mainly of calcium that was deposited by the remains of marine animals). The fragment *a sedimentary rock* is represented in the hypernyms of relations *limestone* → *carbonate sedimentary rock* → *sedimentary rock*; the fragment *consisting mainly of calcium* will be addressed by a meronym relation; finally, *that was deposited by the remains of marine animals* is not mentioned by GBA's definition. The first two parts can be removed without losses. As for the last part, (Encyclopaedia Britannica, 2018) states *limestone has two origins: (1) biogenic precipitation from seawater, the primary agents being lime-secreting organisms and foraminifera; and (2) mechanical transport and deposition of preexisting limestones, forming clastic deposits.* Therefore, the whole WN definition for *limestone* can be disregarded in favor of GBA's¹⁶.

Going through the definitions for these ten synsets so far, one can notice three main aspects covered: the process of forming a rock (e.g.: consolidation, compaction, cementation); the constituents of such rock (e.g.: calcite, aragonite); and the size or aspect of the constituents (e.g.: rounded, >2mm). Focusing on the constituents, we confirm that *rock* is *substance meronym of* 14662574-n (*mineral* : solid homogeneous inorganic substances occurring in nature having a definite chemical composition) in WN. Reflectively, *mineral* is *substance holonym of* *rock*.

Combing through the definitions for the nine terms so far under *rock*, we see that the only minerals referenced are *calcite*, *aragonite* and *dolomite*. All three of them already exist in WN and required only minor changes in the definitions and/or the relations. Essentially the chemical formulas were added to the definitions and the *substance holonym of* relations according to the definitions of the terms we added to WN.

Finally, another set of relations was included: the *domain*

¹⁴Sedimentary material is defined in GBA as *a naturally-occurring material formed at the Earth's surface, consisting of solid particles aggregated together by one or more depositional processes operating within fluid systems (either aqueous or gaseous) to yield granular particles and/or crystalline particles that are aggregated into layers or bodies. The term includes both unconsolidated sediments and sedimentary rocks.*

¹⁵14698000-n (*sedimentary rock* in WN : rock formed from consolidated clay sediments)

¹⁶Limestone definition in GBA is *A carbonate sedimentary rock composed of > 95% calcite (and aragonite) and < 5% dolomite*

of synset - TOPIC and *member of this domain - TOPIC*. Given our topic of choice, all of the terms we added from GBA's lithological terms were associated with *lithology* domain and their constituents with the *mineral* domain.

The *limestone* example shows our approach to map GBA into WN. We included six new and corrected four previously existent synsets definitions, along with their *hypernym of* and *hyponym of* relations. As we analyzed *substance holonym of* and *substance meronym of* relations, we included some of GBA's mineral terms in WN. It is not the scope of this work to cover all of GBA's minerals, but we included the ones mentioned in the rock's definitions.

Following this same approach, we were able to include all of *carbonate sedimentary rock* and *clastic sedimentary rock*, encompassing 27 new synsets with new 79 relations and 9 definitions changes, 15 removed relations and 71 new relations in pre-existing synsets.

These types of sedimentary rocks represent the two main types of oil & gas reservoirs throughout the world. By having them on WN we expect to move one step ahead in NLP for the Oil & Gas domain. We also expect that our time invested in ensuring proper synset relations will improve the performance of word sense disambiguation (WSD) algorithms, specially ones that rely on WN's graph such as UKB (Agirre and Soroa, 2009). At this point we ran the analysis covered in 7.. After the positive results, we carried on with our approach and finished the inclusion and mapping of all GBA lithology material into the WN. With this we expect to move one step further in NLP not only for the Oil & Gas domain but for all geological-related domains, such as Mining, Seismology, and so on.

6. Discussions

The extension of WN raised some relevant points. This section covers such points and explains the reasoning behind the decisions made within the possibilities considered.

A recurring matter regards the multiword expression (MWE) issue. Should we keep and create a synset for an MWE? Or is it enough to have all words individually in the resource? For instance, in WN we have 14698000-n (*sedimentary rock* : rock formed from consolidated clay sediments), but is it a 14696793-n (*rock* : material consisting of the aggregate of minerals like those making up the Earth's crust; "that mountain is solid rock"; "stone is abundant in New England and there are many quarries") that is 02952109-a (*sedimentary* : resembling or containing or formed by the accumulation of sediment; "sedimentary deposits")? Likewise, GBA subdivides *sandstone*, *sand*, *siltstone*, *silt* and *gravel* into *fine*, *medium* and *coarse*, meaning *fine* presents more and smaller grains than *medium* which in turn has more and smaller grains than *coarse*. But GBA sets a specific grain diameter range for *fine sandstone* which is different from the range of *fine siltstone* (respectively 0.063mm to 0.200mm and 0.0020mm to 0.0063mm). Due to this aspect, one possibility would be to adjust existing (or create new) synsets to ensure that *fine*, *medium* and *coarse* retain their relative properties, but the cutoff values (e.g.:0.063mm to 0.200mm) would be lost. In such cases we chose to respect our authoritative source.

Another issue we faced was when layman's knowledge

clashes with technical definitions. For instance, 14698000-n (sedimentary rock : rock formed from consolidated clay sediments): from a technical perspective, clay is an unconsolidated sediment with very small grain, whilst *sedimentary rock* can be formed from several grain sizes, so we replaced WN’s definition with GBA’s. Another example is 14995541-n (sandstone : a sedimentary rock consisting of sand consolidated with some cement (clay or quartz etc.)). Even though WN’s definition was not so far off, it presented *sandstone* as an *hyponym* of 14697485-n (arenaceous rock : a sedimentary rock composed of sand), a term not present in GBA. On the technical side *sand* is a clastic sediment within a certain grain size range, but on the other hand WN defines sand as being silica-based, i.e., the sand commonly found in beaches. This is a common misunderstanding even among technicians. In order to accommodate such divergent points, we merged *arenaceous rock* and *sandstone* synsets, kept the seven synsets *sandstone* was already *hyponym* of and then complemented with GBA’s material.

7. Evaluation

In order to assess the impact of our project, we tested the same NLP pipeline in the same corpus once with the original PWN and once with our extended WN on its intermediary version, i.e. with only *carbonate sedimentary rock* and *clastic sedimentary rock* structures. The results confirmed the value of our approach and justified the inclusion of the remaining GBA’s lithological terms.

The corpus used is one studied by (Rademaker, 2018). It consists of over five thousand sentences, with an average 28 words per sentence. It was built from 1298 publicly available English language geological reports, published by the United States Geological Survey, Geological Survey of Canada and British Geological Survey. The processing was done using Freeling 4.1 (Padró and Stanilovsky, 2012), with the corpus organized in one sentence per file.

The use of our OWN-EN implied in 910 words with different results. Nine had improper Part-of-Speech (PoS) tags and no sense attributed, and for those all PoS and senses were properly attributed with our OWN-EN, but only three to our new synsets - the other six were allocated to previously existing synsets. Such phenomena also happened where the PoS was already correct: of 78 words without allocated synsets, 69 were attributed to previous synsets and only 9 to new synsets. Another 184 words changed synsets within preexisting ones. Finally, there were 639 occurrences of *sandstone* that properly changed from the original WN synset to our previously discussed synset.

One interesting aspect that arises from such numbers is that, *sandstone* apart, most changes were to preexisting synsets. This shows the impact of adding and correcting relations within already existing synsets.

Another relevant case is the change from 13483488-n (formation : natural process that causes something to form; “the formation of gas in the intestine”; “the formation of crystal”; “the formation of pseudopods”) to 09287968-n (formation : (geology) the geological features of the earth) for 59 occurrences of *formation*. Each case was checked, and the switch was judged appropriate for 51 of them. For

the remaining eight cases the original synset was deemed correct.

Conglomerate has fourteen occurrences in the corpus, all of which were previously mapped to 08058937-n (conglomerate : a group of diverse companies under common ownership and run as a single organization) and afterwards were properly mapped to 14863031-n (conglomerate : a composite rock made up of particles of varying size). Each case was individually validated. To illustrate, an example sentence is presented below - clearly it is not about a group of companies, but rather composite rocks.

- (1) On Pliocene and Pleistocene Siwalik Group fluvial sandstones and conglomerates mark the top of the stratigraphic column in the area

8. Conclusion

We were able to expand WordNet from an authoritative source, the Geological Survey of Austria Thesaurus (GBA). The process tackled with evaluating existing synsets for correctness when compared to GBA and creating new synsets otherwise. Such analysis comprehended not only definitions but also the conceptual relations that characterize WordNet.

A three step approach was used. We first used *limestone* as a proof of concept, then all of *carbonate sedimentary rock* and *clastic sedimentary rock*, the main types or reservoir rocks for Oil & Gas. The impact of such extension was evaluated with a corpus containing over five thousand sentences. The results indicated not only the relevance of new synsets added but also the impact conceptual relations changes have on old synsets. Finally, we extended WN to all of GBA’s lithology.

Another product is the mapping between the extended WN synsets and GBA. Because GBA is also mapped to WikiData, BGS, Inspire, GeoSciML and DBpedia, our mapping links such resources as well. This mapping and the extended WN is available at <https://github.com/own-pt/own-en>.

9. Bibliographical References

- Agirre, E. and Soroa, A. (2009). Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL ’09, pages 33–41, USA, 01. Association for Computational Linguistics.
- Amaro, R. and Mendes, S. (2012). Towards merging common and technical lexicon wordnets. In *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon*, pages 147–160, Mumbai, India, December. The COLING 2012 Organizing Committee.
- Antoniak, M., Dalgliesh, J., Verkruyse, M., and Lo, J. (2016). Natural language processing techniques on oil and gas drilling data. In *Intelligent Energy International Conference*, pages 1–6, September.
- Bond, F. and Bond, A. (2019). Geonames wordnet (gnwn): extracting wordnets from geonames. In *Wordnet Conference*, page 387.

- Buitelaar, P. and Sacaleanu, B. (2002). Extending synsets with medical terms. *Proceedings of the First International Conference on Global WordNet*, pages 21–25.
- Encyclopaedia Britannica, T. E. o. (2018). *Limestone*. Encyclopædia Britannica, inc.
- Fellbaum, C. (1998a). *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.
- Christiane Fellbaum, editor. (1998b). *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.
- Isaac, A. and Summers, E. (2009). Skos simple knowledge organization system primer. *Working Group Note, W3C*.
- McCrae, J. P., Rademaker, A., Bond, F., Rudnicka, E., and Fellbaum, C. (2019). English wordnet 2019—an open-source wordnet for english. In *Wordnet Conference*, page 245.
- Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. (1990). Wordnet: An on-line lexical database. *International Journal of Lexicography*, 3:235–244.
- Muniz, H., Chalub, F., Rademaker, A., and de Paiva, V. (2018). Extending wordnet to geological times. In *Global Wordnet Conference 2018*, Singapore, January.
- Padró, L. and Stanilovsky, E. (2012). Freeling 3.0: Towards wider multilinguality. In *LREC2012*.
- Palkowsky, B. (2005). A New Approach to Information Discovery - Geography Really Does Matter. In *SPE Annual Technical Conference*, pages 9–12, Dallas, October.
- Parliament, E. and of the Council. (2007). Inspire directive 2007/2/ec. In *Official Journal of the European Union*, volume 50, page 371–es.
- Peters, W., Sagri, M. T., Tiscornia, D., and Castagnoli, S. (2006). The lois project. In *LREC*, 01.
- Poprat, M., Beisswanger, E., and Hahn, U. (2008). Building a biowordnet by using wordnet’s data formats and wordnet’s software infrastructure: A failure story. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing, SETQA-NLP ’08*, page 31–39, USA. Association for Computational Linguistics.
- Rademaker, A., Tessarollo, A., Pease, A., and Muniz, H. (2019). Extending sumo to geological times. In João Paulo A. Almeida, et al., editors, *Proceedings of the XII Seminar on Ontology Research in Brazil*, volume 2519, pages 70–82, Porto Alegre, RS, September. See <http://ceur-ws.org/Vol-2519/>.
- Rademaker, A. (2018). Challenges for information extraction in the oil and gas domain. In Joel Luís Carbonera, et al., editors, *Proceedings of the XI Seminar on Ontology Research in Brazil (ONTOBRAS)*, São Paulo, Brazil.
- Roventini, A., Antonietta, A., Bertagna, F., Calzolari, N., Jessica, C., Girardi, C., Magnini, B., Marinelli, R., Speranza, M., and Zampolli, A. (2003). Italwordnet: building a large semantic database for the automatic treatment of italian. *Linguistica computazionale : XVIII/XIX, 1998/1999*.
- Sagri, M. T., Tiscornia, D., and Bertagna, F. (2004). Jurwordnet.
- Schoen, E., Smith, R., and Boden, J. (2018). AI Supports Information Discovery and Analysis in an SPE Research Portal. In *SPE Annual Technical Conference and Exhibition*. Society of Petroleum Engineers, September.
- Smith, B. and Fellbaum, C. (2004). Medical wordnet: A new methodology for the construction and validation of information resources for consumer health. In *Proceedings of the 20th International Conference on Computational Linguistics, COLING ’04*, page 371–es, USA. Association for Computational Linguistics.
- van Assem, M., Gangemi, A., and Schreiber, A. (2006). Rdf/owl representation of wordnet. Technical report, World-Wide Web Consortium W3C. <https://www.w3.org/TR/wordnet-rdf/>.
- Vossen, P. (2002). Eurowordnet general document. eurowordnet (le2-4003, le4-8328), part a, final document.

Adding Pronunciation Information to Wordnets

Thierry Declerck^{1,2}, Lenka Bajčetić², Melanie Siegel³

¹German Research Center for Artificial Intelligence, Multilinguality and Language Technology Lab
Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany

²Austrian Centre for Digital Humanities and Cultural Heritage, Austrian Academy of Sciences
Sonnenfelsgasse 19, 1010 Vienna, Austria

³Hochschule Darmstadt, University of Applied Sciences
Max-Planck-Str. 2, 64807 Dieburg, Germany

declerck@dfki.de, Lenka.Bajcetic@oeaw.ac.at, melanie.siegel@hda.de

Abstract

We describe on-going work consisting in adding pronunciation information to wordnets, as such information can indicate specific senses of a word. Many wordnets associate with their senses only a lemma form and a part-of-speech tag. At the same time, we are aware that additional linguistic information can be useful for identifying a specific sense of a wordnet lemma when encountered in a corpus. While work already deals with the addition of grammatical number or grammatical gender information to wordnet lemmas, we are investigating the linking of wordnet lemmas to pronunciation information, adding thus a speech-related modality to wordnets.

Keywords: Wordnet, Pronunciation, OntoLex-Lemon

1. Introduction

Wordnets are well-established lexical resources with a wide range of applications. For more than twenty years they have been elaborately set up and maintained by hand, especially the original Princeton WordNet of English (PWN) (Miller, 1995; Fellbaum, 1998). In recent years, there have been increasing activities in which open wordnets for different languages have been automatically extracted from various resources and enriched with lexical semantics information, building the so-called Open Multilingual Wordnet (OMW) (Bond and Paik, 2012). OMW brings together wordnets in different languages, harmonizing them in a uniform tabular format that lists synsets IDs and the associated lemmas, and linking them to PWN (Bond and Foster, 2013; Bond et al., 2016). Additionally, XML versions of LMF and *lemon* representations¹ of the data are provided.

A starting motivation for our work was to investigate if and how specific Wordnet senses can be restricted to what appears to be morphological variations of a lexical entry. The question touched also the issue on how to encode this information. (Gromann and Declerck, 2019) describe a first experiment done for English, looking at specific Princeton WordNet senses associated with word forms that look like regular plural forms of a lexical entry, but which rather need to be considered as separate lexical entries, due to the specific sense(s) they carry. And PWN is indeed introducing plural forms as “lemmas” in its inventory, when those are related to specific synsets. An example of this is given by the WordNet entry “silks” with the sense of “the brightly colored garments of a jockey; emblematic of the stable”, which is distinct from the synsets associated to the two sin-

gular form entries included in PWN.²

The work described in the present article is an extension of recent experiments done in linking wordnets with additional lexical and morphological information, including grammatical number in the case of PWN (Gromann and Declerck, 2019), grammatical number and grammatical gender in the case of a German lexical semantics resource (Declerck et al., 2019) and of wordnets for Romance languages that are included in OWN (Racioppa and Declerck, 2019). In this context, we note that the Dutch WordNet was from its beginning including full lexical information for a large number of its entries (Vossen et al., 2008; Postma et al., 2016).

In the present work, we investigate the linking of pronunciation information to wordnets, dealing first with the German language. The pronunciation information is extracted from the corresponding German edition of Wiktionary.³

2. Pronunciation as Indicator of Senses

We are aware that different senses of a word, also within a shared part-of-speech category, can be marked by a distinctive pronunciation, like for example for the German substantive “Boot” (in IPA⁴ notation [bu:t]: *boot*) versus “Boot” ([bo:t]: *boat*).⁵ This phenomenon, also called heteronymy, can be relevant for a variety of speech-based ap-

¹LMF stands for “Lexical Markup Framework”, an ISO standard. See (Francopoulo et al., 2006) and <http://www.lexicalmarkupframework.org/> for more details. *lemon* stands for “LEXicon MOdel for ONtologies”. See (McCrae et al., 2012) and <https://lemon-model.net/> for more details.

²This information is retrieved from the PWN Web interface, accessible at <http://wordnetweb.princeton.edu/perl/webwn>.

³See <https://www.wiktionary.org/> and for the German edition <https://de.wiktionary.org/wiki/Wiktionary:Hauptseite>.

⁴IPA stands for “International Phonetic Alphabet”. See <https://www.internationalphoneticassociation.org/content/ipa-chart> for more details.

⁵The pronunciation information is taken from <https://de.wiktionary.org/wiki/Boot>.

plications. Therefore, this type of information should be added to wordnets, so that they can help to disambiguate words in spoken utterances.

We need to make this linking of Wordnet entries to pronunciation information explicit, and for this we are adapting the approach described in (Racioppa and Declerck, 2019), and which is dealing with the linking of Wordnet lemmas to morphological information. We thus again chose the OntoLex-Lemon model (Cimiano et al., 2016)⁶ as the representation formalism, since this model has proven to be able to accommodate both “classical” lexicographic descriptions (McCrae et al., 2017) as well as lexical semantics networks like wordnets (McCrae et al., 2014).

In the next sections, we give first some background description on the extraction of pronunciation information from Wiktionary sources. We continue with a section on OntoLex-Lemon, followed by a section that describes how OntoLex-Lemon supports the linking of lemmas in wordnets resources to pronunciation information.

3. Extracting Pronunciation Data from Wiktionary

It has been shown that the access and use of Wiktionary can be helpful in a series of Natural Language Processing (NLP) applications. (Kirov et al., 2016), for example, describe work to extract and standardize data contained in Wiktionary and to make it available for a range of NLP tasks, while the authors focus on extracting and normalizing a huge number of inflectional paradigms across a large selection of languages. This effort contributed to the creation of the UniMorph data (<http://unimorph.org/>). The UniMorph project was focusing on (scraping) the HTML representation of Wiktionary (mostly the English version, but also looking at other language editions). (Metheniti and Neumann, 2018) and (Metheniti and Neumann, 2020) describe a related approach, but making use of a combination of the HTML pages and the underlying XML dump of the English edition of Wiktionary, which is covering also 4,050 other languages, some of them with a very low number of entries.⁷ The English edition of Wiktionary has of today a number of 6,262,000 pages, whereas 734,130 pages are dealing with English words.

BabelNet⁸ is also integrating Wiktionary data,⁹ with a focus on sense information, in order to support, among others, word sense disambiguation and tasks dealing with word similarity and sense clustering (Camacho-Collados et al., 2016).

Many language specific editions of Wiktionary contain also pronunciation information, mostly encoded with the help of

⁶See also <https://www.w3.org/2016/05/ontolex/> for more details.

⁷A possibly tentative list of entries in the different languages contained in the English Wiktionary is given here: <https://en.wiktionary.org/wiki/Special:Statistics?action=raw>.

⁸See (Navigli and Ponzetto, 2010) and <https://babelnet.org/>.

⁹As far as we are aware of, BabelNet integrates only the English edition of Wiktionary, but includes all the languages covered by this edition.

the IPA notation. (Jouvet et al., 2011) show that pronunciation information encoded in (the French edition of) Wiktionary can be “used efficiently for building a pronunciation lexicon for a speech transcription system”. (Schlippe et al., 2010) assess the quality of pronunciation information in Wiktionary for four languages (English, French, German, and Spanish) and come to satisfying results, especially in the case of French, when it comes to the evaluation of the coverage and also to the impact on automatic speech recognition (ASR) systems, especially in the case of Spanish. Those already older studies comforted us in the opinion that extracting pronunciation information from Wiktionary can deliver a relevant source of data for our experiment consisting in equipping wordnets with pronunciation information.

4. Extracting Pronunciation Information from the German Edition of Wiktionary

We display in Figure 1 below as an example the pronunciation information for the German substantive “Januar” (*january*) as represented in the XML dump of the German edition of Wiktionary.¹⁰ As the reader can see, the

```

{{Aussprache}}
:{{IPA}} {{Lautschrift|'janua:ɐ}}
:{{Hörbeispiele}} {{Audio|De-Januar.ogg}}

```

Figure 1: The Wiktionary markup encoding of the pronunciation of the German word “Januar” (*january*).

information on the pronunciation is encoded in the wiki markup language, and the element names are in German (“Aussprache” standing for *pronunciation*, “Lautschrift” for *phonetic script* and “Hörbeispiele” for *audio samples*). This means that for every language edition of Wiktionary a specific script has to be written for extracting the desired information. Also the use of the wiki markup is not consistent across language editions, so that the scripts have also to be adapted for dealing with the various templates in use in the different language editions.


A first version of our extraction program allowed us to detect a (provisional, as the extraction script can still be improved) list of 150 German substantives that have two or more pronunciations.¹¹ We are extending this list to other categories, also looking for words belonging to more than one category, as for example “modern” (adjective, [moˈdɛʁn], *modern*) versus “modern” (verb, [moːdɛʁn], *moulder*). But this cross-categories extension is less relevant, as wordnets would anyway introduce different lemmas for a word belonging to distinct categories.

An example of a German substantive having two different pronunciations is “Vollzug”, with the stress put either at the

¹⁰XML dumps of the various editions of Wiktionary are available at <https://dumps.wikimedia.org/backup-index.html>.


¹¹In parallel, we are extracting a list of German substantives that have different genders (502 entries detected) or different plural forms (440 entries detected), each with specific senses.

beginning or at the end of the word, as shown in Figure 2 and Figure 3, which are displaying screen shots from the Wiktionary page, and where the reader can see the meanings (encoded as the values of the key word “Bedeutungen”) associated with the distinct pronunciations.¹²

Aussprache:
 IPA: [ˈfɔl,tʁu:k]
 Hörbeispiele:  [Vollzug](#) (Info)

Bedeutungen:
 [1] (U- oder S-)Bahn-Garnitur, welche die übliche Länge aufweist
 [2] Güterzug, der mit Fracht beladen ist

Figure 2: The German word “Vollzug” in Wiktionary, with the meanings of *train set* and *charged freight train*.

Aussprache:
 IPA: [fɔl'tʁu:k]
 Hörbeispiele:  [Vollzug](#) (Info)
 Reime: -u:k

Bedeutungen:
 [1] Umsetzung in die Tat, das Ausführen
 [2] *kurz für:* [Strafvollzug](#)
 [3] Einrichtung, in der [Verurteilte](#) ihre [Freiheitsstrafe](#) [absitzen](#)

Figure 3: The German word “Vollzug” in Wiktionary, with the meanings of *execution* [1] and *enforcement, penal system, prison* [2],[3].

Our internal representation for the pronunciation information, together with the associated meanings, extracted from the XML dump of Wiktionary is displayed in Figure 4. This is the type of data to be linked to synsets for German, making use for this of the OntoLex-Lemon representation model.

```
Vollzug
['fɔl'tʁu:k']
["Umsetzung in die Tat, das Ausführen
 \n:2 ''kurz für:'' Strafvollzug
 \n:3 Einrichtung, in der Verurteilter|
 Verurteilte ihre Freiheitsstrafe
 absitzen\n\n{"}
['fɔl,tʁu:k']
["(U- oder S-)Bahn-Garnitur, welche die
 übliche Länge aufweist
 \n:2 Güterzug, der mit Fracht beladen
 ist\n\n{']
```

Figure 4: Our internal representation of the extracted pronunciation information, with the associated meanings, from Wiktionary for the word “Vollzug”.

5. OntoLex-Lemon

OntoLex-Lemon is a further development of the “Lexicon Model for Ontologies” (*lemon*) (McCrae et al., 2012). Both

¹²<https://de.wiktionary.org/wiki/Vollzug>.

lemon and the OntoLex-Lemon model, which is resulting from a W3C Community Group,¹³ were originally developed with the aim to provide a rich linguistic grounding for ontologies, meaning that the natural language expressions used in the labels, definitions or comments of ontology elements are equipped with an extensive linguistic description.¹⁴ This rich linguistic grounding includes the representation of morphological and syntactic properties of lexical entries as well as the syntax-semantics interface, i.e. the meaning of these lexical entries with respect to an ontology or to specialized vocabularies.

The main organizing unit for those linguistic descriptions is the *LexicalEntry* class, which enables the representation of morphological patterns for each entry (a multi word expression, a word or an affix). The connection of a lexical entry to an ontological entity is marked mainly by the denotes property or is mediated by the *LexicalSense* or the *LexicalConcept* classes, as this is represented in Figure 6, which displays the core module of the model.

A major difference between *lemon* and OntoLex-Lemon is that the latter includes an explicit way to encode conceptual hierarchies, using the SKOS¹⁵ standard. As can be seen in Figure 6, lexical entries can be linked via the *ontolex:evokes* property to such SKOS concepts, which can represent Wordnet synsets. This structure is paralleling the relation between lexical entries and ontological resources, which is implemented either directly by the *ontolex:reference* property or mediated by the instances of the *ontolex:LexicalSense* class.

As can be seen in Figure 6, there is a property called *ontolex:phoneticRep* which is introduced for the class *ontolex:Form*. This property is used in the model for representing the pronunciation information, which is thus encoded at the level of morphological forms and not at the level of lexical entries, as this is shown in Figure 5 for the example entry “privacy”:

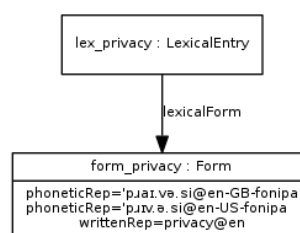


Figure 5: The graphical representation of the place of the “*ontolex:phoneticRep*” property in the OntoLex-Lemon model. Taken from <https://www.w3.org/2016/05/ontolex/#forms>

¹³See <https://www.w3.org/2016/05/ontolex/>

¹⁴See (McCrae et al., 2012) and (Cimiano et al., 2016).

¹⁵SKOS stands for “Simple Knowledge Organization System”. SKOS provides “a model for expressing the basic structure and content of concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies, folksonomies, and other similar types of controlled vocabulary” (<https://www.w3.org/TR/skos-primer/>).

More recently, OntoLex-Lemon has been used also as a de-facto standard in the field of digital lexicography and is being applied for example in the European infrastructure project ELEXIS (European Lexicographic Infrastructure).¹⁶

Our present goal is to integrate synsets, lemmas, morphological and pronunciation descriptions in the extended ontological framework specified by OntoLex-Lemon. Updating also past work on mapping some wordnets to the former *lemon* model (McCrae et al., 2014). This work was done following the guidelines¹⁷ for mapping Global WordNet formats onto *lemon*-based RDF.¹⁸

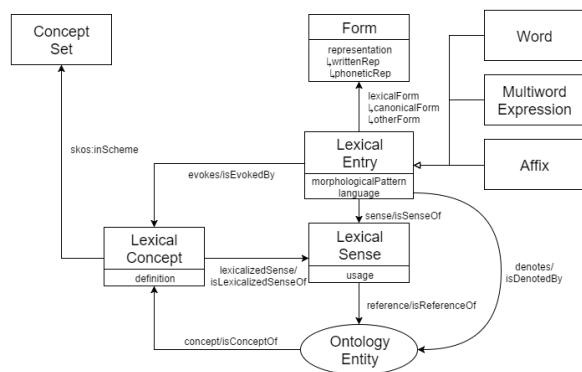


Figure 6: The core module of OntoLex-Lemon. Graphic taken from <https://www.w3.org/2016/05/ontolex/>.

6. The integrated Encoding in OntoLex-Lemon

We display in code listing 1 the (still tentative) way we can express the phonetic restriction for a sense of an OdeNet¹⁹ concept that points to the word “Vollzug”.

Listing 1: The OntoLex-Lemon representation of the OdeNet synset for the concept associated with *Vollzug* pointing to all listed entries senses and a corresponding form

```

: synset_odenet -2345-n
  rdf:type ontolex:LexicalConcept ;
  wn: ili ili:i41311 ;
  skos:inScheme :OdeNet ;
  ontolex:isEvokedBy :entry_w10755 ;
  ontolex:isEvokedBy :entry_w11251 ;
  ontolex:isEvokedBy :entry_w11252 ;

```

¹⁶See <http://www.elex.is/> for more detail.

¹⁷See <https://globalwordnet.github.io/schemas/#rdf>.

¹⁸RDF stands for “Resource Description Framework”. See <https://www.w3.org/RDF/> for more details.

¹⁹“OdeNet” stands for “Open-de-WordNet”. See (Declerck et al., 2019) for more info on OdeNet, a lexical semantics resource for German. The original resource (still under development) can be downloaded here: <https://github.com/hdaSprachtechnologie/odenet>.

```

ontolex:isEvokedBy :entry_w11253 ;
ontolex:isEvokedBy :entry_w11254 ;
ontolex:isEvokedBy :entry_w11255 ;
ontolex:isEvokedBy :entry_w11256 ;
ontolex:isEvokedBy :entry_w11257 ;
ontolex:isEvokedBy :entry_w11258 ;
ontolex:isEvokedBy :entry_w11259 ;
ontolex:isEvokedBy :entry_w11260 ;
ontolex:isEvokedBy :entry_w7091 ;
ontolex:lexicalizedSense
  :sense_w10755_2345-n ;
ontolex:lexicalizedSense
  :sense_w11251_2345-n ;
ontolex:lexicalizedSense
  :sense_w11252_2345-n ;
ontolex:lexicalizedSense
  :sense_w11253_2345-n ;
ontolex:lexicalizedSense
  :sense_w11254_2345-n ;
ontolex:lexicalizedSense
  :sense_w11255_2345-n ;
ontolex:lexicalizedSense
  :sense_w11256_2345-n ;
ontolex:lexicalizedSense
  :sense_w11257_2345-n ;
ontolex:lexicalizedSense
  :sense_w11258_2345-n ;
ontolex:lexicalizedSense
  :sense_w11259_2345-n ;
ontolex:lexicalizedSense
  :sense_w11260_2345-n ;
ontolex:lexicalizedSense
  :sense_w7091_2345-n ;

```

```

:entry_w11258
  rdf:type ontolex:Word ;
  wn:partOfSpeech wn:noun ;
  ontolex:canonicalForm :form_w11258 ;
  ontolex:evokes :synset_odenet -2345-n ;
  ontolex:evokes :synset_odenet -3815-n ;
  ontolex:sense :sense_w11258_2345-n ;
  ontolex:sense :sense_w11258_3815-n ;

```

```

:sense_w11258_2345-n
  rdf:type ontolex:LexicalSense ;
  ontolex:isLexicalizedSenseOf
    :synset_odenet -2345-n ;
  ontolex:isSenseOf :entry_w11258 ;
  lexicog:restrictedTo
    :form_w11258.Restriction_2 .

```

```

:form_w11258
  rdf:type ontolex:Form ;
  ontolex:writtenRep "Vollzug"@de ;

```

```

:form_w11258.Restriction_2
  rdf:type ontolex:Form ;
  ontolex:phoneticRep "vollZUG"@de ;

```

The most important part of this encoding is the property `lexicog:restrictedTo` added to the one `Lex-`

icalSense that is relevant in our case. This property has been defined in a recent extension to the core module of OntoLex-Lemon: the “lexicog” module, which has been developed for covering specific aspects of Lexicography.²⁰ We then introduce a specific object called “form_w11258_Restriction_2”, which encodes for the :form_w11258 the special case of the second pronunciation for “Vollzug”, as displayed in Figure 3.²¹ This way we can not only add pronunciation information to wordnets, but also express the restriction that a specific meaning is dependant on a specific pronunciation.

7. Conclusion

We described work in progress consisting in adding pronunciation information to wordnets, as this information can be very relevant in making wordnets usable for sense disambiguation in speech applications. Using for this purpose the OntoLex-Lemon model allows us not only to encode this linking from original wordnets to pronunciation information extracted from Wiktionary dictionaries, but this supports also the possibility to express restrictions on senses, stating that a specific sense can be only selected in case a specific pronunciation is given.

8. Acknowledgements

This paper is supported by the Horizon 2020 research and innovation programme with the projects Prêt-à-LLOD (grant agreement no. 825182) and ELEXIS (grant agreement no. 731015). It is also partially based upon work from COST Action CA18209 - NexusLinguarum “European network for Web-centred linguistic data science”. We also thank the anonymous reviewers for their helpful comments.

9. Bibliographical References

- Bond, F. and Foster, R. (2013). Linking and extending an open multilingual wordnet. In *ACL (1)*, pages 1352–1362.
- Bond, F. and Paik, K. (2012). A survey of wordnets and their licenses. *Small*, 8(4):5.
- Bond, F., Vossen, P., McCrae, J. P., and Fellbaum, C. (2016). Cili: the collaborative interlingual index. In *Proceedings of the Global WordNet Conference*, volume 2016.
- Bosque-Gil, J., Lonke, D., Gracia, J., and Kernerman, I. (2019). Validating the OntoLex-lemon lexicography module with K Dictionaries’ multilingual data. In *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference.*, pages 726–746, Brno, Czech Republic, October. Lexical Computing CZ s.r.o.,
- Camacho-Collados, J., Pilehvar, M. T., and Navigli, R. (2016). Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240:36–64.

²⁰See <https://www.w3.org/2019/09/lexicog/> and (Bosque-Gil et al., 2019) for more details.

²¹We mark with capital letters the fact that the stress is on the second part of the word.

- Cimiano, P., McCrae, J. P., and Buitelaar, P. (2016). Lexicon Model for Ontologies: Community Report.
- Declerck, T., Siegel, M., and Gromann, D. (2019). Ontolex-lemon as a possible bridge between wordnets and full lexical descriptions. In Christiane Fellbaum, et al., editors, *Proceedings of the Tenth Global Wordnet Conference*, pages 264–271, wyb. Stanisława Wyspiańskiego 27 50-370 Wrocław Poland, 7. Oficyna Wydawnicza Politechniki Wrocławskiej, Oficyna Wydawnicza Politechniki Wrocławskiej.
- Christiane Fellbaum, editor. (1998). *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA.
- Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M., and Soria, C. (2006). Lexical markup framework (lmf). In *International Conference on Language Resources and Evaluation-LREC 2006*, page 5.
- Gromann, D. and Declerck, T. (2019). Towards the detection and formal representation of semantic shifts in inflectional morphology. In Maria Eskevich, et al., editors, *2nd Conference on Language, Data and Knowledge (LDK)*, volume 70 of *OpenAccess Series in Informatics (OASICs)*, pages 21:1–21:15. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 5.
- Jouvet, D., Fohr, D., and Illina, I. (2011). Building a Pronunciation Lexicon for a Speech Transcription System from Wiktionary Pronunciations only. In *XIV International Conference “Speech and Computer” (SPECOM’2011)*, Kazan, Russia, September.
- Kirov, C., Sylak-Glassman, J., Que, R., and Yarowsky, D. (2016). Very-large scale parsing and normalization of wiktionary morphological paradigms. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).
- McCrae, J., de Cea, G. A., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., and Wunner, T. (2012). Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, 46(6):701–709.
- McCrae, J. P., Fellbaum, C., and Cimiano, P. (2014). Publishing and linking wordnet using lemon and rdf. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics*.
- McCrae, J. P., Buitelaar, P., and Cimiano, P. (2017). The OntoLex-Lemon Model: Development and Applications. In Iztok Kosem, et al., editors, *Proceedings of eLex 2017*, pages 587–597. INT, Trojína and Lexical Computing, Lexical Computing CZ s.r.o., 9.
- Metheniti, E. and Neumann, G. (2018). Wikinflection: Massive semi-supervised generation of multilingual inflectional corpus from wiktionary. In *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018)*, Linköping Electronic Conference Proceedings. Linköping University Electronic Press, Linköpings universitet, 12.

- Metheniti, E. and Neumann, G. (2020). Wikinflection corpus: A (better) multilingual, morpheme-annotated inflectional corpus. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*. LREC.
- Miller, G. A. (1995). Wordnet: A lexical database for english. *COMMUNICATIONS OF THE ACM*, 38:39–41.
- Navigli, R. and Ponzetto, S. P. (2010). BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden, July. Association for Computational Linguistics.
- Postma, M., van Miltenburg, E., Segers, R., Schoen, A., and Vossen, P. (2016). Open dutch wordnet. In *Proceedings of the Eight Global Wordnet Conference*, Bucharest, Romania.
- Racioppa, S. and Declerck, T. (2019). Enriching open multilingual wordnets with morphological features. In Raffaella Bernardi, et al., editors, *Proceedings of the Sixth Italian Conference on Computational Linguistics*. CEUR, 10.
- Schlippe, T., Ochs, S., and Schultz, T. (2010). Wiktionary as a source for automatic pronunciation extraction. In *11th Annual Conference of the International Speech Communication Association, Makuhari, Japan*. Interspeech 2010.
- Vossen, P., Maks, E., Segers, R., and van der Vliet, H. (2008). Integrating lexical units, synsets and ontology in the cornetto database. In European Language Resources Association (ELRA), editor, *Proceedings of LREC 2008, Marrakech*.

Author Index

Agnihotri, Shikhar, 7

Alkorta, Jon, 20

Alvez, Javier, 1

Bajcetic, Lenka, 39

Bond, Francis, 14

Chakraverty, Shampa, 7

Declerck, Thierry, 39

Garg, Apar, 7

Gonzalez-Dios, Itziar, 1, 20

Marciniak, Jacek, 25

McCrae, John Philip, 14

Mohapatra, Soumya, 7

Rademaker, Alexandre, 14, 33

Rigau, German, 1

Rudnicka, Ewa, 14

Shah, Praveen, 7

Siegel, Melanie, 39

Tessarollo, Alexandre, 33