International Journal of Computational Linguistics and Applications, Vol. 6, No. 2, 2015, pp. 61–82
Received 23/01/2015, Accepted 27/02/2015, Final 06/04/2015.
ISSN 0976-0962, http://ijcla.bahripublications.com

knoWitiary: A Machine Readable Incarnation of Wiktionary

VIVI NASTASE
CARLO STRAPPARAVA
FBK-irst, Trento, Italy

ABSTRACT

knoWitiary *is a resource that presents a reorganized version of Wiktionary's information in machine readable format. Wiktionary contains a plethora of information about words, including sense definitions, etymology, translations, derived terms and anagrams. Similar work to the one reported here goes one step further than extracting information from Wiktionary: mapping it onto WordNet – NLP community's de facto gold standard. Lexical and relation overlap shows that Wiktionary provides different types of information compared to WordNet, which implies that much is discarded when doing a mapping. We make a case here for making space for "pure" resources alongside mapped ones, to preserve the unique information that idiosyncratic resources such as Wiktionary provide, which may open up new avenues to explore for tasks that require varied and "unorthodox" information about words.*

1. INTRODUCTION

knoWitiary is a network of words and senses obtained exclusively from Wiktionary. We compare it with one of the most used resources in NLP – WordNet. Because WordNet has been in use for more than two decades now, we know its strengths and weaknesses. Work on resources that are similar to it – deal with words, their senses and relations between them –

often falls back onto WordNet, to take advantage of its manually built (and therefore, we consider it) gold standard hierarchy and inventory. By "fall back" we mean that instead of building a new resource from scratch, the work concentrates on adding to this resource, enriching it with new words, senses or relations between them. This may seem appropriate from several points of view – the "backbone" of the newly proposed resource is not called into question since it is WordNet itself; adoption of the new resource can be faster, since it adheres to WordNet's format and conventions. There may not be only advantages in adopting WordNet as the core of a new resource. First, criticisms of WordNet itself – its occasionally too fine-grained sense distinctions, differences in the level of detail in hierarchies governed by different "top senses" – would apply to the new resources. But the most important disadvantage of using WordNet as a scaffolding to build upon is that the structure and inventory it imposes enforces compromises in the kind of knowledge that is being added to it: mapping a dictionary or Wikipedia onto WordNet requires a mapping at the level of senses, and there does not exist a one-to-one mapping between entries in these resources and WordNet. This leads to further cuts in the amount and type of information that was extracted and could have been made available. Thus, the mapping causes a compromise, with loss of information and structure. These losses are not quantified in such work, as the more positive aspects – the mapping process and the enrichment of the "base" resource, usually WordNet – are emphasized (e.g. [1, 2]).

We proceed to build a stand-alone resource by formalizing Wiktionary as comprehensively as possible. Wiktionary[1] is a very rich dictionary, that presents in semi-structured format a plethora of information about words: senses and glosses, phonology, derivations, word relations within a language and across languages. It also contains "unorthodox" relations, such as ANAGRAMS[2] and ETYMOLOGY. This treasure trove of

[1] http://en.wiktionary.org/wiki

[2] Incidentally, note that *knoWitiary* is an anagram of *Wiktionary*

interconnected information is unprecedented, and could bring new exploration avenues for established NLP tasks, and the needed spark for novel creative language tasks. We compare this resource with WordNet – the English 3.1[3] and the Italian versions[4]. The comparison shows a large amount of novel information, only part of which can be imported when performing a mapping [3].

Versions of Wiktionary formatted for machine consumption already exist, including a freely available Java library for processing the Wiktionary dump (JWKTL). Each of these has some piece missing. We will review these versions in Section 2. In Section 3 we describe knoWitiary and its general statistics in terms of multi-lingual lexica and relations. The comparison with English and Italian wordnets is described in Section 4, and we wrap up with a brief overview of tasks that could be aided or made possible by having a resource with the kind of varied information that Wiktionary contains in Section 5.

2. RELATED WORK

WordNet [4] has for many years been the lexical resource used in NLP research. Built by psycholinguists and lexicographers, its structure relies on the notion of synset – a set of one or more synonyms that expresses a "unit" of meaning, linked through several types of lexico-semantic relations with other synsets: semantic (e.g. HYPERNYM, HYPONYM; three types of MERONYM and HOLONYM; SYNONYM, ANTONYM); lexical (DERIVED FROM, PERTAINYMS, PARTICIPLE OF); domain / member of domain. An index serves to map word forms under four parts of speech (adjective, adverb, noun, verb) onto synsets, and data files include the relations between synsets.

WordNet has provided the gold standard in word senses (used as reference for multiple word sense disambiguation exercises within Sens-/Sem-Eval) and ontology (used as a

[3] http://wordnet.princeton.edu/wordnet/download/
[4] http://multiwordnet.fbk.eu/english/home.php

reference for ontological relation extractions) for the general English language. Having such a strong backbone, rather than build something new from scratch, there have been efforts to produce enhanced wordnets – with coarser word senses [5], in different languages [6], with sentiment annotation [7, 8], with domains [9] with more relations [10-12].

Wiktionary[5] is an online collaborative dictionary, companion to **Wikipedia**[6], which provides a collaborative wiki platform for the building of dictionaries in multiple languages. Reflecting the varied knowledge of the contributors, Wiktionary contains comprehensive information about words. For a given *word form* we may find each applicable part of speech and language, alternative spellings, (layered) senses and definitions, phonology, etymology, synonyms, hypernyms, hyponyms, derived and related terms, translations, anagrams, and images.

Just like Wikipedia, Wiktionary is semi-structured: it has sections for each of the main types of information it provides, while the information within the section can be structured or not. For *derived forms* for example we find the related entries listed, while etymological information appears in a free-form paragraph, but which contains structured word information, and often regular patterns to express etymological links.

Various types of information extracted from Wiktionary have been exploited successfully for a variety of NLP tasks, such as semantic relatedness measures [13], cross-language image retrieval [14], named entity recognition [15], synonymy mining [16], cross-language text categorization [17]. Such results have shown that Wiktionary is a desirable resource, and its availability in machine-readable format would be an asset to NLP applications.

JWKTL[7] is an API to Wiktionary, available as a free Java library. It processes a Wiktionary dump and populates a database with information for English, Russian and German. The library

[5] http://en.wiktionary.org
[6] http://en.wikipedia.org
[7] https://www.ukp.tu-darmstadt.de/software/jwktl/

provides very fast processing of the Wiktionary dump, and varied and flexible methods to access the formalized information. Because the entries in the dictionary are not all structured, some information is lost in conversion. In particular, the etymological information is presented as a string (the paragraph as it was on the wiki page), without further formalization. With respect to the other information contained therein, and bar some parsing errors on either side, this resource and knoWitiary and roughly equivalent.

de Melo [18] describes *Etymological WordNet*, built from etymological links mined from the *Etymology* and *Derived from* sections, and also definitions. This resource is part of a larger repository, described in [19], built based on a few assumptions that define and prescribe the definition and design of universal multilingual knowledge bases. The concrete work done towards achieving such a comprehensive resource uses WordNet as the base, and it builds upon it by adding mono-lingual (in particular language family information under the corresponding synset for *language*) and multi-lingual entries (based on translations from Wiktionary), and novel – etymological – links. At the time of writing, this resource was not available for analysis and comparison.

Mappings Since WordNet, Wiktionary and Wikipedia do not subsume one another, there has been effort in various combinations of mapping between the three [2]. Any such mapping imposes the adoption of one resource as the "base", onto which the others are mapped. Because of its good reputation and ubiquity in the field, this "base" is (usually) WordNet.

Miller and Gurevych [2], Gurevych et al. [20] present mappings between WordNet, Wiktionary, Wikipedia and other sources, and include overviews of previous work on mapping between various combinations of such resources. One interesting thing to note is that while considerable effort is made to make, evaluate and report the mapping at the level of nodes, it is not clear what happens with the relations from other sources, or with the un-mapped nodes. Through the mapping, the aim is to show how much the resource (onto which the mapping is done) gets

enriched, but we are missing the final picture – what does the final resource contain. By not investigating the un-mapped portions, we don't know how much of the potential of the other resources remains untapped.

3. FROM WIKTIONARY TO KNOWITIARY

The entries in Wiktionary are semi-structured. There are sections for definitions, etymology, pronunciation, and for each part of speech that may apply, there are derived terms and translations. Different etymologies for the same word are presented in different sections – thus allowing, if necessary, the distinction between homonymy and polysemy. In terms of word senses, there are both coarse and fine-grained distinctions, where a coarse sense may have several sub-senses. This type of information is illustrated through the entry for the word *form* in Figure 1. As apparent from the figure, we note that Wiktionary is organized by word forms. If the same word form appears in another language than English, it appears within the same Wiktionary page, with the same type of information as for English.[8]

[8] We remind the reader that we are processing here only the English Wiktionary (the version from 10.04.2014) where all information apart from the words themselves (if needed) is given in English. It covers however word forms in multiple languages. Wiktionaries for other languages exist, but were not included in the resource described here.

Figure 1. form *in Wiktionary*

We process the structured portions of the page, and for each section of interest extract the available information. We extract first the words and their possible senses and sub-senses with the associated definitions (Figure 2). 29 of the languages represented have each more than 10,000 entries, under 16 parts of speech. Table 1 shows part of the lexicon and relation statistics, for some of the most populated languages.

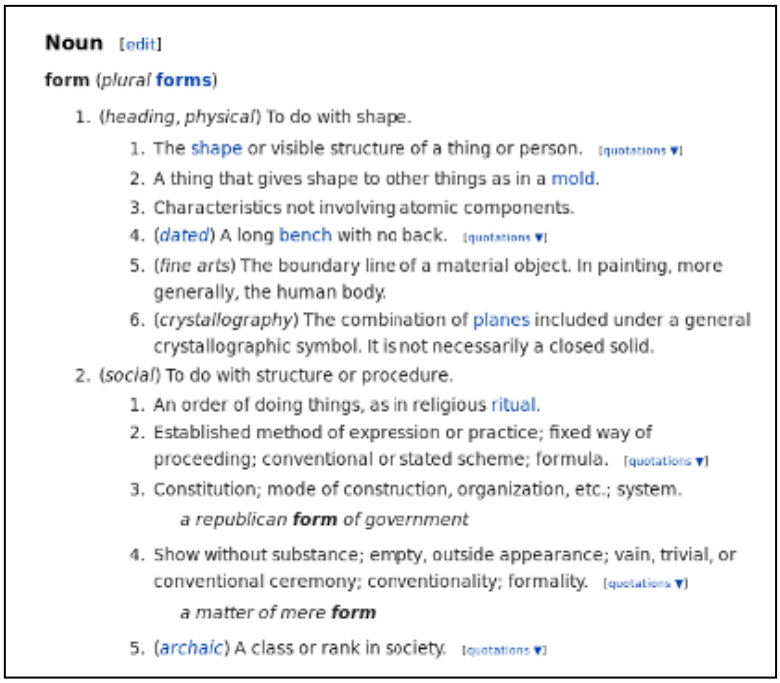

Figure 2. *Definitions*

Table 1. *Selected lexicon and relation statistics in knoWitiary*

Language	# entries	# senses	# subsenses	# relations
English	581,586	702,575	706,466	710,396
French	291,291	126,142	126,202	84,181
German	169,118	231,692	231,727	307,872
Italian	529,630	270,341	270,394	671,689
Latin	661,642	634,588	635,982	589,674
29 most frequent	3,605,984	3,610,320	3,616,321	3,307,981

For each form there is varied information, including related terms, synonyms, antonyms, derived terms, as illustrated in Figure 3. An overview of the relations extracted from these sections (with statistics covering the 29 most represented languages) is included in the top part of Table 2.

Figure 3. *Related words*

Table 2. *Relation statistics in knoWitiary*

General relations		
Relation	**freq.**	**Example**
ACRONYM OF	572	NATO / North Atlantic Treaty Organization
ALTERNATIVE FORMS	91,781	encyclopedia / encyclopaedia
ANAGRAMS	442,422	dictionary / indicatory
ANTONYMS	47,090	free / bound
COMPOUNDS	16,969	live (adj) / live broadcast
CONJUGATION OF	991,759	it:abbreviate / it:abbreviare
DERIVED TERMS	305,339	book (noun) / bookworm
DESCENDANTS	44,069	la:dictionarium (noun) / en:dictionary
HOLONYMS	856	nucleotide / dezoxyribonucleic acid
HYPERNYMS	46,905	mouse / rodent
HYPONYMS	46,908	deer / buck
MERONYMS	856	conjunction / conjunct
RELATED	550,731	lexicography / lexicon
SEE ALSO	106,146	dictionary / vocabulary
SYNONYMS	360,779	book (noun) / tome
total	3,053,182	
Etymological relations (direct)		
Relation	**freq.**	**Example**
ABBRV	365	en:bot / en:robot
BORROWING	2,782	fr:sandwich / en:sandwich
COGNATE	4	en:meal / nl:moal
COGNATE COMPOUND	5	nl:Aalderik / goh:adal:noble + goh:rihhi:ruler
COMPOUND	54,685	la:dictionarius / la:dictio:speaking: + la:-arium:room, place:
CONFIX	8,504	en:morphology / en:morpho
ETYM	188,448	en:dictionary / la:dictionarium
ETYMTWIN	6	en:word / en:verb
total	254,799	

The etymology of words is presented in a "free form" paragraph, which however uses a rather consistent lexicon and expressions to present the information. Figure 4 shows an example. To extract the etymological chain, we parse the *Etymology* section using regular expressions. Statistics on the direct links extracted from these sections are shown in the second half of Table 2.

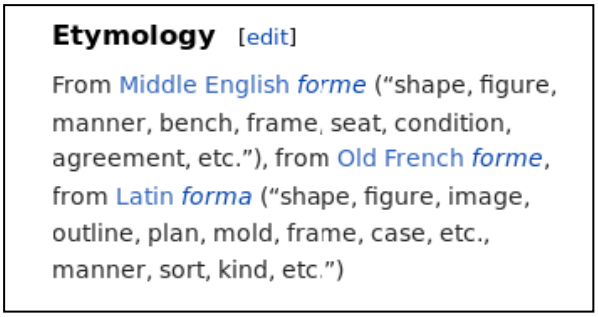

Figure 4. *Etymology in Wiktionary*

Because the Wiktionary entries are "stand alone" (edited individually, and not necessarily coordinated with existing entries), during the construction process we add the symmetrical (ANTONYM, SYNONYM, RELATED) and opposite relations (MERONYM/HOLONYM, HYPERNYM/HYPONYM) to those explicitly given. From the translation section we extract translations. Translations are grouped by sense, and as a link to the word senses a brief version of the sense definition is given, as can be seen in Figure 5.

Figure 5. *Translations by sense*

There are 122,571 entries for 74,384 English words, which have 1,494,527translations. There are multiple entries per word because the translations are at the sense level. There are 2,384 languages represented, 36 of which have more than 10,000 occurrences as translations.

It can be argued that Wiktionary is not a trustworthy source of lexical knowledge, because of its open nature and its collaborative and (probably) non-expert pedigree. We discuss below the issues of lexical material. That Wiktionary contains useful information is evidenced by its contribution to NLP tasks, as explained in Section 2.

Lexical material: root forms There are entries in Wiktionary that do not appear in dictionaries such as Merriam-Webster, Collins, Oxford – e.g. widespreadness, tiltability, hypotheticality. Professionally built dictionaries go through a rigorous process of selecting the lexical material. Word usage is surveyed through selected sources of text, and novel words are

"quarantined" until they become established in the language[9]. This "quarantine" was on the order of years, but has become shorter to keep up with the productivity of language speakers and the high rate of information exchange and spread facilitated by the web and numerous electronic social platforms. The three words mentioned above do not (as yet) qualify for inclusion in such dictionaries, but each has thousands of hits on the web, and appear in various sources including scientific or technical publications. Having an up-to-date inventory of language, even if some entries are destined to fall by the wayside, is useful, whether dealing with contemporary texts, or dealing with older texts that contain words that in the meantime have disappeared from language. There may be situations when Wiktionary contributors may add a new, made-up word that they like. It is not likely that entries like this affect the rest of the resource: if a word that is included is never used, it is not an issue.

Lexical material: inflected forms For inflective languages, such as Italian, the inflected forms may appear as separate entries, whereas proper dictionaries include only the root form and the applicable inflectional rules. These decisions were necessary for paper dictionaries for reasons of space. In an electronic version it is not necessary to censor inflected forms. The statistics in Table 1 & 2 show that the Italian entries cover a high number of inflected forms. We argue that this is neither a problem, nor a negative aspect of the resource. From a practical point of view, inflected entries in the machine readable version of the resource makes recognition easier and faster by simple string match, without need for lemmatization or stemming. There is also another aspect related to the inflected forms, which explains why they are included in the Wiktionary at all: as mentioned before, Wiktionary is organized by word forms. The same word form may appear in different languages, whether as a root or inflected form. All but 16,200 of the forms that have an entry in Italian have entries in other languages as well. For example, the

[9] http://www.oxforddictionaries.com/words/how-do-you-decide-whether-a-new-word-should-be-included-in-an-oxford-dictionary

inflected form *minute* – the plural feminine version of the adjective *minuto* (tiny) – has entries in English, French and Latin as well, some which are inflections (for Italian and Latin) some of which are root forms (for English and French). This parallel between forms in different language is itself an interesting bit of information which is captured by the resource.

4. COMPARISON WITH WORDNET

Since WordNet is the most commonly used ontology in NLP, the question of how the new resource compares comes naturally. For the English and Italian portions of the extracted resource we perform comparison with the corresponding wordnets in terms of lexicon – entries and senses – and relations.

The purpose of the comparison is to quantify both the portions that can be mapped, but most interestingly, those that cannot. The fact that much information cannot be mapped supports the idea that on the field of lexical resources there should be space for "pure" resources other than those manually built by experts, and onto which mappings are done.

4.1. *Lexical comparison*

Table 3 provides numbers for comparison in terms of senses between WordNet and knoWitiary. An apparent advantage of working with Wiktionary is the fact that it has a two-level structure – senses and subsenses, which would allow access at varying levels of granularity, depending on the task. The table contains statistics on the number of forms and senses for each of the parts-of-speech represented in WordNet, and also sense/subsense information.

Table 3. *Sense statistics for words in WordNet and knoWitiary*

POS	**WordNet (EN)**			**knoWitiary (EN)**				
	# forms		# senses	# forms		# senses	# subsenses	
adj	21,499	30,070	(1.398)	91,218	110,179	(1.208)	110,478	(1.211)
adv	4,475 5,	592	(1.25)	15,251	17,397	(1.141)	17,435	(1.143)
noun	117,953	146,512	(1.242)	378,206	457,147	(1.208)	459,870	(1.215)
verb	11,540	25,061	(2.171)	92,589	116,914	(1.263)	117,759	(1.272)
total	155,467	207,235	(1.33)	577,264	701637	(1.215)	705,542	(1.222)
POS	**WordNet (IT)**			**knoWitiary (IT)**				
	# forms		# senses	# forms		# senses	# subsenses	
adj	5,074	6,452	(1.271)	63,215	67,399	(1.066)	67,415	(1.066)
adv	1,634	2,250	(1.376)	3,996	4,915	(1.23)	4,915	(1.23)
noun	34,935	49,219	(1.408)	95,059	108,743	(1.144)	108,762	(1.144)
verb	4,969	9,875	(1.987)	366,138	374,301	(1.022)	374,303	(1.022)
total	46,612	67796	(1.454)	528,408	555,358	(1.051)	555,395	(1.051)

Table 4 includes statistics on the frequency of the different parts-of-speech in Wiktionary, and the overlap with WordNet (the English 3.1 and Italian versions) for the POS classes represented in WordNet: adjectives, adverbs, nouns and verbs. The high number of word forms for verbs in Italian is caused by the inclusion of inflections. These numerous inflected entries provide a de-facto inflectional derivational resource for Italian. The statistics shown here were obtained from one English Wiktionary dump[10]. Meyer et. al [1] show an overview of Wiktionary coverage (2012), and for three languages (English, German, Russian) extensive comparison between the lexicon from the three Wiktionary versions and the corresponding wordnets, and the coverage of several word lists representing the basic vocabulary of each of the three languages. Meyer et al. [1]'s analysis shows that Wiktionary's coverage of the basic vocabulary is very high, thus supporting its use as a lexical reference resource.

Wiktionary's coverage of WordNet's vocabulary is high, but the majority of Wiktionary entries are not included in WordNet. Even for shared vocabulary, Meyer et al. [3] show that even with a high accuracy sense mapping (estimated on a set of 2,423 pairs of WordNet-Wiktionary senses), more than 370,000 Wiktionary

[10] Version from 10.04.2014: enwiktionary-20141004-pages-articles.xml

senses remain unmapped (on the Wiktionary version used). This shows that when including only the mapped vocabulary and senses, the majority of the potential information to be added is in fact discarded.

Table 4. *Lexical overlap with WordNet*

POS	Overlap	coverage rel. to WN	freq. In EN WordNet	freq. In knoWitiary (EN)
adjective	15,924	74.07%	21,499	91,218
adverb	3,837	85.74%	4,475	15,252
article			–	15
cardinal numeral			–	90
conjunction			–	226
determiner			–	122
interjection			–	2,055
noun	51,283	43.48%	117,953	378,212
numeral			–	200
participle			–	3
preposition			–	501
pronoun			–	441
suffix			–	644
verb	9,837	85.24%	11,540	92,590
total	80,881	52.02%	155,467	581,586

Lexical overlap with Italian WordNet

POS	Overlap	coverage rel. to WN	freq. In IT WordNet	freq. In knoWitiary (IT)
adjective	4,119	81.18%	5,074	63,215
adverb	1,386	84.82%	1,634	3,996
article			–	12
conjunction			–	158
interjection			–	202
noun	21,273	60.89%	34,935	95,050
numeral			–	1
participle			–	4
preposition			–	312
pronoun			–	210
suffix			–	322
verb	4,260	85.73%	4,969	366,139
total	31,038	66.59%	46,612	529,630

4.2. *Relation comparison*

Relations such as SYNONYMS, HYPERNYMS, HYPONYMS, ANTONYMS appear in both WordNet and Wiktionary (in

Wiktionary they appear as sections, which we then formalized as relations), but other relations are particular to one or the other of the resources. The way they are encoded is also different. In Wiktionary, most of the relations presented here (except the etymological ones) appear as section headers, with the related terms presented as a list. In WordNet most relations are between synsets, with the relation of a word belonging to a certain synset explicit through index files. The synonymy relation is implicit between words belonging to the same synset. Relations between synsets can apply to all or specific elements of the synset, and this is signaled within the data files. When computing the number of instances for each WordNet relation this will be taken into account to obtain the correct number of relations in the resource. For the comparison all relation extracted from Wiktionary are named, and from WordNet the 10 most frequent relations are considered separately[11], with all the others grouped under OTHER.

The low overlap in terms of relations show that Wiktionary covers qualitatively different information than WordNet. In the previous work involving Wiktionary and mapping it onto WordNet, the focus is on the mapping of word senses. Had the relation between the mapped word senses been also added, the enrichment in this respect would have been small compared to the original size of the resource: 77,548 pairs of English words (not senses, though) appear in Wiktionary and are connected through a relation, and appear but are not connected in WordNet. Compared with WordNet 3.1's original size (1 million+ relations), the increase is small. The real richness would have come from additional entries that could not be mapped, and their relations.

[11] In WordNet there are three types of meronym and holonym relations. While it is a useful distinction, for the statistics we use only the coarser MERONYM/HOLONYM.

Table 5. *Relation overlap between knoWitiary and English WordNet (3.1)*

Relation overlap: English WordNet → knoWitiary mapping					
knoWit. relation	overlap	overlap with homonym. relation	highest overlap (*WN* rel.)	freq. in WordNet	freq. in knoWitiary
ABBRV	–	–	–	–	179
ACRONYM OF	–	–	–	–	502
ALTERNATIVE FORMS	3,009	–	SYNONYMS (2,882)	–	53,229
ANAGRAMS	239	–	SYNONYMS (192)	–	114,743
ANTONYMS	2,296	2,167	ANTONYMS (2,167)	7,983	18,792
BORROWING	–	–	–	–	805
COGNATE	–	–	–	–	4
COMPOUND	–	–	–	–	11,341
COMPOUNDS	–	–	–	–	8
CONJUGATION OF	–	–	–	–	3
CONFIX	–	–	–	–	2,964
DERIVED TERMS	9,728	3,968	DERIVED TERMS (3,968)	74,680	126,317
DESCENDANTS	8	–	DERIVED TERMS (5)	–	273
ETYM	–	–	–	–	47,684
ETYMTWIN	–	–	–	–	6
HOLONYMS	67	45	HOLONYMS (45)	103,246	377
HYPERNYMS	766	597	HYPERNYMS (597)	364,600	6,661
HYPONYMS	766	597	HYPONYMS (597)	364,600	6,664
MERONYMS	67	45	MERONYMS (45)	103,246	377
RELATED	13,336	–	DERIVED TERMS (7,577)	137,209	111,995
SEE ALSO	3,781	23	SYNONYMS (1,161)	4,732	57,827
SYNONYMS	27,609	14,570	SYNONYMS (14,570)	157,394	149,645
total	61,672	22,012		1,180,481	710,396
Relation overlap: knoWitiary → English WordNet mapping					
WN relation	overlap	overlap with homonym. relation	highest overlap (*knoWit* rel.)	freq. in WordNet	freq. in knoWitiary
ANTONYMS	2,683	2,167	ANTONYMS (2,167)	7,983	18,792
ATTRIBUTE	406	–	RELATED (236)	3,418	–
DERIVED FROM/PERTAINYM	1,230	–	RELATED (1,108)	8,074	–
DERIVED TERMS	12,177	3,968	DERIVED TERMS (3,968)	74,680	126,317
HOLONYMS	881	45	SYNONYMS (346)	103,246	337
HYPERNYMS	6,184	597	SYNONYMS (4,081)	364,600	6,661
HYPONYMS	10,580	597	SYNONYMS (4,081)	364,600	6,664
MERONYMS	953	45	SYNONYMS (346)	103,246	377
OTHER	4,097	–	SYNONYMS (3,284)	137,209	111,995
SEE ALSO	546	23	SYNONYMS (473)	4,732	57,827
SYNONYMS	21,952	14,570	SYNONYMS (14,570)	157,394	149,645
total	61,672	22,012		1,191,973	478,655

Table 6. *Overlap with the Italian WordNet for the Italian entries in knoWitiary*

Relation overlap: English WordNet → knoWitiary mapping					
knoWit. relation	overlap	overlap with homonym. relation	highest overlap	freq. in WordNet	freq. in knoWitiary
ABBRV	–	–	–	–	5
ACRONYM OF	–	–	–	–	1
ALTERNATIVE FORMS	205	–	SYNONYMS (172)	–	1,598
ANAGRAMS	16	–	SYNONYMS (10)	–	194,225
ANTONYMS	24	2	ATTR./HYPON./ HYPERN (6)	22	3,017
BORROWING	–	–	–	–	114
COMPOUND	–	–	–	–	284
CONJUGATION OF	–	–	–	–	294,688
CONFIX	–	–	–	–	2,998
DERIVED TERMS	329	–	HYPONYMS (182)	–	14,045
DESCENDANTS	2	–	HYPON./HYPERN (1)	–	125
ETYM	–	–	–	–	7,664
HYPERNYMS	9	8	HYPERNYMS (8)	116,318	22
HYPONYMS	9	8	HYPONYMS (8)	116,318	22
RELATED	2,060	–	SYNONYMS (714)	6,909	118,669
SEE ALSO	288	–	SYNONYMS (113)	–	3,410
SYNONYMS	5,852	3,650	SYNONYMS (3,650)	53,153	30,802
total	8,794	3,668		292,720	671,689
Relation overlap: knoWitiary → English WordNet mapping					
WN relation	overlap	overlap with homonym. relation	highest overlap	freq. in WordNet	freq. in knoWitiary
ANTONYMS	2	2	–	22	3,017
ATTRIBUTE	149	–	RELATED (126)	3,372	–
HOLONYMS	143	–	RELATED (96)	8,429	–
HYPERNYMS	1,593	8	SYNONYMS (998)	116,318	22
HYPONYMS	1,767	8	SYNONYMS (998)	116,318	22
MERONYMS	144	–	RELATED (96)	8,429	–
OTHER	263	–	SYNONYMS (170)	6,909	118,669
SYNONYMS	4,733	3,650	SYNONYMS (3,650)	53,153	30,802
total	8,794	3,668		311,851	152,532

5. NOVEL PERSPECTIVES ON NLP TASKS

knoWitiary contains much lexical information that is novel, and combines pieces of information that have previously not been accessible from a single resource.

Etymology, in particular, has proved its usefulness in bridging different languages in a cross-lingual text categorization task [17]. Cross-lingual text categorization consists in categorizing documents in a target language L_t using a model built through supervised learning on a labeled dataset in source language L_s. The task is even more difficult when the datasets in the two languages are not parallel (there is a 1:1 mapping

between the texts contained in the two datasets, one being the translation of the other), but rather consist of comparable corpora (i.e. documents on the same topics, such as sports, economy). Etymological relations provide a layer of shared word-ancestors that connect the two languages, thus allowing the model to capture text-category associations at this shared linguistic level. Having translations also available would allow this bridge to be further enriched, thus leading to better shared models across the languages.

Etymological information is also crucial in studying the evolution of language during different epochs. It could be possible to study why some words evolve rapidly through time while others stay the same, often with an identical meaning in many different languages. We could verify hypotheses such as: the more often a word is used, the less likely it is to mutate. A similar observation was made relative to verbs, where those that are most commonly used have irregular forms [21].

Instances of compounding and word derivation, in particular in those situations where the resulting term is not compositional (or not any longer) – e.g. $breakfast$ are instances of language creativity that can be further studied to find how such term have been generated, and thus endow a machine with similar capabilities. Measuring the semantic distance between a term and its etymological children could show how metaphors are coined, and what kind of word relations have been used to make this creative jump. The connected nature of Wiktionary would allow for such investigations.

Other creative language tasks would benefit from a rich lexical resource. For example [22] propose a computational approach to generate neologisms consisting of homophonic puns and metaphors based on the category of the service to be named and the properties to be underlined. This kind of task is very challenging from a lexical knowledge point of view, because it requires a combination of semantic, phonetic, lexical and morphological knowledge to automatize the process.

6. CONCLUSIONS

The work presented here had two motivations: (i) to obtain a coherent and consistent lexical resource that contains as much information as possible about words and their relations, (ii) to measure what could be gain and what would be lost by forcing a mapping of such a resource onto another structure. To obtain the lexical resource we processed Wiktionary, the on-line collaboratively built dictionary covering a treasure trove of entries and relations in numerous languages. To measure this against other resources used in NLP, we choose WordNet, since it is the most frequently used, and also the base for mapping other resources, including those based on Wiktionary itself. We have explored both the lexical and relation overlap, which shows that Wiktionary provides a different kind of information than WordNet does. Mapping it onto WordNet would mean discarding such unique information, with unknown impact on both the tasks that use the mapped version, and on tasks that are never attempted because the mapped resource lacks the needed information/links.

REFERENCES

1. Meyer, C. M. & Gurevych, I. 2012. Wiktionary: A new rival for expert-built lexicons? exploring the possibilities of collaborative lexicography. In S. Granger & M. Paquot (Eds.), *Electronic Lexicography* (pp. 259-291). Oxford: Oxford University Press.
2. Miller, T. & Gurevych, I. 2014. WordNet-Wikipedia-Wiktionary: Construction of a threeway alignment. In proceedings of the *9th Language Resources and Evaluation Conference (LREC 2014)*.
3. Meyer, C. M. & Gurevych, I. 2011. What psycholinguists know about Chemistry: Aligning Wiktionary and WordNet for increased domain coverage. In proceedings of the *5th International Joint Conference on Natural Language Processing* (pp. 883-892).
4. Fellbaum, C. (ed.) 1998. *WordNet: An Electronic Lexical Database*. Cambridge, Mass: MIT Press.
5. Mihalcea, R. & Moldovan, D. 2001. Automatic generation of a coarse grained WordNet. In proceedings of *NAACL Workshop on WordNet and Other Lexical Resources*, Pittsburgh, PA, Jun.

6. Vossen, P. (ed.) 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer, Dordrecht, The Netherlands.
7. Strapparava, C., Gliozzo, A. & Giuliano, C. 2004. Pattern abstraction and term similarity for word sense disambiguation. In proceedings of the *3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL-3) at ACL-04* (pp. 229-234), Barcelona, Spain, 25-26 Jul.
8. Baccianella, S., Esuli, A. & Sebastiani, F. 2010. SentWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In proceedings of the *7th International Conference on Language Resources and Evaluation* (pp. 2200-2204), La Valetta, Malta, 17-23 May.
9. Magnini, B., Strapparava, C., Pezzulo, G. & Gliozzo, A. 2002. The role of domain information in word sense disambiguation. *Natural Language Engineering*, 8, 359-373.
10. Snow, R., Jurafsky, D. & Ng, A. Y. 2006. Semantic taxonomy induction from heterogeneous evidence. In proceedings of the *21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics* (pp. 801-808), Sydney, Australia, 17-21 Jul.
11. Boyd-Graber, J., Fellbaum, C., Osherson, D. & Schapire, R. 2006. Adding dense, weighted connections to WordNet. In proceedings of the *Third Global WordNet Meeting*, Jeju Island, Korea, Jan.
12. Ruiz-Casado, M., Alfonseca, E. & Castells, P. 2005. Automatic extraction of semantic relationships for WordNet by means of pattern learning from Wikipedia. In proc. of the *10th International Conference on Applications of Natural Language to Information Systems* (pp. 67-79).
13. Zesch, T., M uller, C. & Gurevych, I. 2008. Using Wiktionary for computing semantic relatedness. In proceedings of the *23rd Conference on the Advancement of Artificial Intelligence* (pp. 861-867), Chicago, Ill., 13-17 Jul.
14. Etzioni, O., Reiter, K., Sonderland, S. & Sammer, M. 2007. Lexical translation with application to image search on the web. In proceedings of the *Machine Translation Summit XI*, Copenhagen, Denmark.
15. Richman, A. E. & Schone, P. 2008. Mining wiki resources for multilingual named entity recognition. In proceedings of the *46th Annual Meeting of the Association for Computational Linguistics:*

Human Language Technologies (pp. 1-9), Columbus, Ohio, 15-20 Jun.
16. Navarro, E., Sajous, F., Gaume, B., Prevot, L., ShuKai, H., Tzu-Yi, K., Magistry, P. & Chu-Ren, H. 2009. Wiktionary and nlp: Improving synonymy networks. In proceedings of the *2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources* (pp. 19-27).
17. Nastase, V. & Strapparava, C. 2013. Bridging languages through etymology: The case of cross language text categorization. In proceedings of the *51st Annual Meeting of the Association for Computational Linguistics* (pp. 651-659), Sofia, Bulgaria.
18. de Melo, G. 2014. Etymological Wordnet: Tracing the history of words. In proceedings of the *9th Language Resources and Evaluation Conference (LREC 2014)*.
19. de Melo, G. & Weikum, G. 2010 .Towards universal multilingual knowledge bases. In *Principles, Construction, and Applications of Multilingual Wordnets*. *Proceedings of the 5th Global WordNet Conference (GWC 2010)* (pp. 149-156), New Delhi, India.
20. Gurevych, I., Eckle-Kohler, J., Hartmann, S., Matuschek, M., Meyer, C. M. & Nghiem, T. D. 2013. Uby – a large-scale lexical-semantic resource. In *Book of Abstracts of the 23rd Meeting of Computational Linguistics in the Netherlands: CLIN 2013* (p. 81)
21. Pinker, S. 1999. *Words and Rules*. Canada: Harper Collins.
22. Özbal, G. & Strapparava, C. 2012. A computational approach to automatize creative naming. In proceedings of the *50th annual meeting of the Association of Comptational Linguistics (ACL-2012)* (pp. 703-711), Jeju Island, Korea.

VIVI NASTASE
FBK-IRST, TRENTO, ITALY.
E-MAIL: <NASTASE@FBK.EU>

CARLO STRAPPARAVA
FBK-IRST, TRENTO, ITALY.
E-MAI: <STRAPPAG@FBK.EU>