

An approach to measuring and annotating the confidence of Wiktionary translations

Antonio J. Roa-Valverde¹ · Salvador Sanchez-Alonso² · Miguel-Angel Sicilia² · Dieter Fensel¹

Published online: 6 February 2017
© Springer Science+Business Media Dordrecht 2017

Abstract Wiktionary is an online collaborative project based on the same principle than Wikipedia, where users can create, edit and delete entries containing lexical information. While the open nature of Wiktionary is the reason for its fast growth, it has also brought a problem: how reliable is the lexical information contained in every article? If we are planing to use Wiktionary translations as source content to accomplish a certain use case, we need to be able to answer this question and extract measures of their confidence. In this paper we present our work on assessing the quality of Wiktionary translations by introducing confidence metrics. Additionally, we describe our effort to share Wiktionary translations and the associated confidence values as linked data.

Keywords Linguistics · Linked data · Ranking · Random walks

✉ Antonio J. Roa-Valverde
antonio.roa@sti2.at

Salvador Sanchez-Alonso
salvador.sanchez@uah.es

Miguel-Angel Sicilia
msicilia@uah.es

Dieter Fensel
dieter.fensel@sti2.at

¹ Semantic Technology Institute, University of Innsbruck, Innsbruck, Austria

² University of Alcalá, Madrid, Spain

1 Introduction

The Web is turning into a data oriented platform. Collaborative initiatives like Linked Open Data (LOD) rely on accepted standards and provide with a set of best practices to promote the sharing and consumption of data at large scale. A growing portion of this data is populated by linguistic information, which tackles the description of lexicons and their usage. An important resource within this scope is Wiktionary,¹ which can be seen as the leading data source containing lexical information nowadays. Wiktionary is an online collaborative project based on the principle of the “Wisdom of Crowd” that tries to build an open multilingual dictionary available for everybody. Since its inception in 2002, Wiktionary has grown considerably and, therefore, caught the attention of many researchers. Several attempts have tried to compare the usability of Wiktionary with traditional expert-edited lexicographical efforts (Fuentes-Olivera 2009; Meyer and Gurevych 2012). Others have relied on reusing the data provided by Wiktionary for accomplishing certain information retrieval (IR) and natural language processing (NLP) tasks (Müller and Gurevych 2009; Zesch et al. 2008; Navarro et al. 2009). Additional approaches have focused on aligning Wiktionary with other available resources (Matuschek et al. 2013; Miller and Gurevych 2014).

In this paper we exploit the multilingual dimension of Wiktionary. We perform a quantitative analysis of the existing translations in order to measure their level of reliability and guarantee a minimum of quality during their consumption. We rely our analysis on the use of ranking approaches like random walks, which have shown to provide successful results in IR and other related scenarios. A part of our research focuses on the study of existing formats and mechanisms to exchange linguistic data. We use the gained expertise to design an ontological model in order to cope with the interoperability issues associated to our scenario. We use this model to share the generated data as part of the public open data cloud.

The paper is organized in two main parts. The first part focuses on the description of our approach for associating confidence values to Wiktionary translations. In Sect. 2 we describe the multilingual structure of Wiktionary. Section 3 introduces our approach. We describe the formalisms of our algorithm in Sect. 3.1 and a quantitative evaluation in Sect. 3.2. The second part, which starts in Sect. 4, deals with LOD approaches for modeling and sharing linguistic data. In Sect. 4.1 we describe the *lemon* vocabulary. The different efforts for handling multilingual data are described in Sect. 4.2. Section 4.3 describes our approach for adding confidence annotations as an extension to the *lemon* vocabulary. In Sect. 4.4 we show the resulting dataset after representing the generated data from the evaluation with the proposed vocabulary. Finally, Sect. 5 concludes this work and enumerates possible extensions.

¹ <http://wiktionary.org>.

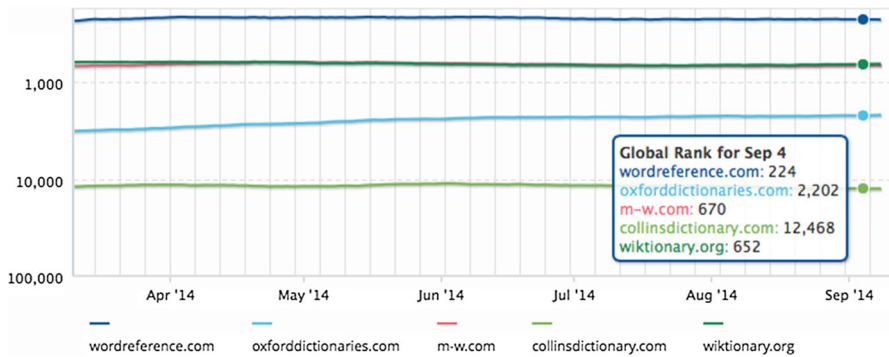


Fig. 1 Dictionary traffic metric comparison by Alexa

2 Wiktionary

The task of compiling lexical data to build dictionaries has been traditionally related to lexicographical experts belonging to well-known educational institutions. Independently of the format chosen to distribute the dictionary (printed vs. digital, offline vs. online), the creation process is characterized for being time-consuming and tedious. The social aspect introduced by the Web 2.0 pushed forward the collaborative nature of this task by embracing wiki-based initiatives to allow any Internet user to add contributions on this area. Wiktionary is a product born out of this initiative. Being a relatively young project in comparison to established dictionary providers, its online presence can give an overview of its traction among the community. Figure 1 shows a comparison of the traffic metrics of Wiktionary and four other known dictionary providers.

The advantages of Wiktionary in comparison to other expert-built resources rely on the following properties:

- *Collaborative* the easy handling of the underlying wiki-based system allows any Web user to contribute. Due to the growing community and its contribution, the data keeps up to date with potential changes in the language. Additionally, there is a group of editors that take care of reviewing the changes and avoid the loss of data due to vandalism.
- *Rich coverage* Wiktionary offers an article for each lexical word, containing diverse information like definitions, part of speech, etymology, senses, lexico-semantic relationships (homonyms, synonyms, antonyms, hyponyms, hyperonyms) and translations. This information acquires a high value in the case of small languages, due to the difficulties to find resources targeting them.

- *Open data* the content offered by Wiktionary can be reused following the Creative Commons CC-BY-SA 3.0² and GNU Free Documentation³ licenses. This makes Wiktionary particularly attractive to app developers willing to build ad-hoc functionality on top of its content without dealing with subscription fees.
- *Multilingual* Wiktionary supports different languages that are organized in separated language editions, each with a different community of supporters and contributors. For instance, the Spanish version of Wiktionary contains information about lexical words in Spanish (in this case, the wiki-UI is rendered to the user also in Spanish). Language editions not only contain information about words within its own language, but also foreign words. In this way, for example, the Spanish Wiktionary contains information about English words that are described in Spanish. The idea behind this feature is to allow the description of foreign words in one's native language.

The available content in the different languages is interlinked by *translations links* and *inter-wiki links*. The former are links within the same language edition reflecting words with equivalent meaning in two languages. E.g., “bank” (English) <http://en.wiktionary.org/wiki/bank> and “banco” (Spanish) <http://en.wiktionary.org/wiki/banco#Spanish>. The latter are links to the same word in a different language edition. E.g., “bank” (English) <http://en.wiktionary.org/wiki/bank> and “banco” (Spanish) <http://es.wiktionary.org/wiki/banco>.

The current mechanism implemented in Wiktionary for adding new translations to a certain word is the following: (a) users navigate through the word's senses until finding the required one within a certain language edition. (b) In the *translations* section of the chosen sense the user must make sure that the translation she is willing to add does not exist for the target language. If it does not, then she can use the form shown at the bottom of the list of available translations to enter the details if using the Web UI (Fig. 2) or just add the scripting code as described in the public guideline⁴ if using the wiki editor. (c) The added translation is reviewed by the responsible editor of the language edition and made available to the public or rejected and removed from the resource.

At the time of writing Wiktionary supports 171 languages and provides more than 21 million articles.⁵

Even though these properties have shown to be leading Wiktionary towards the right direction, the project is still a resource under construction and presents data deficiencies, which make users doubt about the reliability of the information. This lack is deeper for some languages than others, specially when dealing with translations, which has been remarked in public critics to Wiktionary.⁶ In the following, we describe our attempt to provide Wiktionary translations with

² http://en.wiktionary.org/wiki/Wiktionary:Text_of_Creative_Commons_Attribution-ShareAlike_3.0_Unported_License.

³ http://en.wiktionary.org/wiki/Wiktionary:GNU_Free_Documentation_License.

⁴ <http://en.wiktionary.org/wiki/Wiktionary:Translations>.

⁵ <http://meta.wikimedia.org/wiki/Wiktionary>.

⁶ <http://mastersofmedia.hum.uva.nl/2008/09/20/wiktionary-and-the-limitations-of-collaborative-sites/>.

branch office of such an institution [hide ▲]

Select targeted languages

<ul style="list-style-type: none"> • Arabic: مصرف <i>m</i> (másríf), بنك <i>m</i> (bank) • Armenian: բանկ (bank) • Asturian: bancu <i>m</i> • Basque: banketxe • Breton: ti-bank ^(br) <i>m</i> • Bulgarian: банков клон (bankov klon) • Catalan: banc <i>m</i> • Chinese: <ul style="list-style-type: none"> Mandarin: 銀行 ^(zh), 银行 ^(zh) (yínháng) • Crimean Tatar: bank • Czech: banka ^(cs) • Danish: bank ^(da), bankfilial • Dutch: bank ^(nl) <i>f</i> • Esperanto: banko • Estonian: pank • Finnish: pankki ^(fi) • French: banque ^(fr) <i>f</i> • German: Bank ^(de) <i>f</i> • Greek: τράπεζα ^(el) <i>f</i> (trápeza), υποκατάστημα ^(el) <i>n</i> (ypokatástima) • Hebrew: בנק ^(he) <i>m</i> (banq) 	<ul style="list-style-type: none"> • Icelandic: banki ^(is) <i>m</i>, bankaútibú <i>n</i> • Indonesian: bank ^(id) • Interlingua: banca • Italian: banca ^(it) <i>f</i> • Japanese: 銀行 ^(ja) (ginkyō), 銀行 (banku) • Jèrriais: banque <i>f</i> • Korean: 은행 ^(ko) (eunhaeng) (銀行 ^(ko)) • Macedonian: банка <i>f</i> (bánka) • Norwegian: bank ^(no) <i>m</i> • Polish: bank ^(pl) <i>m</i> • Portuguese: banco ^(pt) <i>m</i> • Russian: банк ^(ru) <i>m</i> (bank) • Serbo-Croatian: <ul style="list-style-type: none"> Cyrillic: банка <i>f</i> Roman: banka ^(sh) <i>f</i> • Slovene: banka ^(sl) <i>f</i> • Spanish: banco ^(es) <i>m</i>, sucursal <i>f</i> • Swedish: bank ^(sv) <i>c</i>, bankkontor <i>n</i> • Volapük: bank ^(vo) • West Frisian: bank
--	--

• Add translation :
Preview translation [More](#)
 Script code: (e.g. Cyril for Cyrillic, Latn for Latin)

Fig. 2 Wiktionary—“Add translation” form

confidence annotations in order to help the user to get an idea of the reliability of the data.

3 Confidence analytics

Several publications in the literature have tried to address the problem of measuring the reliability of Wikipedia articles. Most of the approaches rely on the analysis of the article’s history to extract some measures that can be correlated with the grade of quality. In this direction, Lih (2004) suggested a correlation of the quality of an article with the amount of editors and the number of revisions. Lim et al. (2006) proposed a ranking of Wikipedia articles based on three different metrics: the length of the article, the number of revisions and the reputation of the editors, which is dependent on the amount of previous edits. Other approaches focus on analyzing the text content of the articles. Blumenstock (2008) used metrics like the amount of words, characters and sentences within an article. These metrics were used to classify the articles as featured and non-featured. Authors achieved an accuracy of 97%, demonstrating that the number of words is the best metric to perform such classification.

Contrary to the case of Wikipedia, assessing the quality of Wiktionary content is an area that has been less explored. Fuertes-Olivera (2009) evaluate the adequacy of

Wiktionary from a lexicographical point of view regarding to some social needs, for instance, the use case of learning business-related vocabulary. In this work, authors manually perform a qualitative analysis of the lexical information available for terms related to the business topic. This analysis inspects the structure and the kind of lexical content (definitions, part of speech, phonetics, use examples, translations, etc.) used to describe each entry. The analysis focuses only on the English–Spanish scenario, but authors remark the English-centered nature of Wiktionary, i.e., the amount of content in the English Wiktionary is several grades of magnitude more extensive than in other language editions. Meyer and Gurevych (2012) perform another qualitative analysis for the English, German and Russian editions of Wiktionary and compare these resources with the work done by professional lexicographers. Weale et al. (2009) and Sajous et al. (2013) address the problem of detecting new synonyms for Wiktionary entries. Both works rely on the link structure of Wiktionary and the use of random walks to measure the grade of synonymity among the different candidates. These two approaches are at certain stage the closest to our work we could find in the bibliography. Despite these analyses, to the best of our knowledge, no quantitative approaches have been published addressing the reliability of Wiktionary translations. In the following we describe our contribution.

3.1 Translation-confidence algorithm

The approach we present does not target the quality of the individual Wiktionary articles as a whole, like it has been done in previous research efforts on Wikipedia, but it focuses only on translations. Unlike history-based approaches, that mostly use text mining, we rely our computation on link analysis.

Previously, we described how translations are organized in Wiktionary following two kinds of links, namely, *translation links* and *inter-wiki links*. As already discussed, these links define the structure of the multilingual dimension of Wiktionary. We base our approach on the following hypothesis: *the way users contribute by adding new translations or modifying existing ones could have an impact on how the multilingual structure is organized*. We argue that user contributions and the link structure formed by translations are strongly related, being possible to extract some measures that reflect the reliability of translations. We think the wisdom of the crowd can be used to reflect the quality of the translations available in Wiktionary at certain stage. If Wiktionary users add links to a certain word, there are more chances for this word to be correct and for the translation pairs to be relevant. In different words, we believe there is a relation between popularity of words and translation quality.

Note that the heuristic we propose in this work does not decide itself whether a translation pair is correct or not, it rather gives a confidence score as additional information to the user that could support her with this decision.

Our approach uses PageRank (Brin and Page 1998) to estimate the quality of the translation pairs. PageRank is an algorithm originally conceived for ranking Web documents. Considering the structure of the Web as a directed graph, where the nodes are the Web documents and the edges are the different links among Web

documents, the idea behind PageRank is that of assigning more relevance to those documents with a bigger amount of incoming links. Authors consider this value as a measure of the attention towards a certain document. From another perspective, this relevance represents the probability that a random surfer ends up in a Web document after following a certain amount of links. The relevance of a Web document is higher when it is referenced by important documents. This fact implies a recursive definition of rank and therefore, the rank of a web document is a function of the ranks of the Web documents referencing it.

Applying the previous concept, we build an input graph relying only on the use of *inter-wiki links* and discarding the *translation links*. The reason for this is that *translation links* are strongly dependent of each language edition and they do not always contain rich information about the translations, i.e., there are many entries for which the translation link points to a to-be-done Wiktionary page.⁷

There are other considerations to take into account that have a clear impact on the resulting ranking. First, due to the varying amount of contributions for each supported language, the different language editions must be split when building the input graph. This means that only those links within the same language pair will be considered. This approach will avoid penalizing those translations which are popular in some language pairs, but are not available for others. Second, it is necessary to organize the translations in what we have called an *Isolated Semantic Graph* (ISG). An ISG is a directed graph where the nodes are existing entries in a language edition and the edges are *inter-wiki links* from a single source language edition to a single target language edition. A couple of nodes joined by an *inter-wiki link* form a translation pair. The key about ISGs is that they contain only translations for those words that are related semantically, i.e., there are *inter-wiki links* at n-hops connecting the words. Figure 3 shows the ISG associated to the English word “able” and targeting the Spanish language edition.

3.1.1 Notation

We use the following notation to present our algorithm. Let WLE be the set of all available language editions of Wiktionary. We denote every single language edition by WLE_{lang} , where $lang$ is the ISO code corresponding to the particular language. For instance, WLE_{en} represents the English language edition of Wiktionary. Every language edition contains a set of words in the same language, $\{w_1^{en}, w_2^{en}, \dots, w_n^{en}\}$.

We define the following function:

$$f_{source \rightarrow target}(w_i^{source}) = \{w_1^{target}, \dots, w_m^{target}\}$$

For a given word w_i^{source} in WLE_{source} , f returns the set of words belonging to WLE_{target} that are connected with w_i^{source} through an *inter-wiki link*.

⁷ An example of this can be found for the German translation of the word “banco” in the Spanish language edition (<http://es.wiktionary.org/wiki/banco>). The available translations contain the German translation “Bank”, however it points to an empty page (<http://es.wiktionary.org/w/index.php?title=Bank&action=edit&redlink=1>), showing that the term does not exist in the Spanish edition. The *translation link* is well created in the opposite direction, i.e., from German to Spanish.

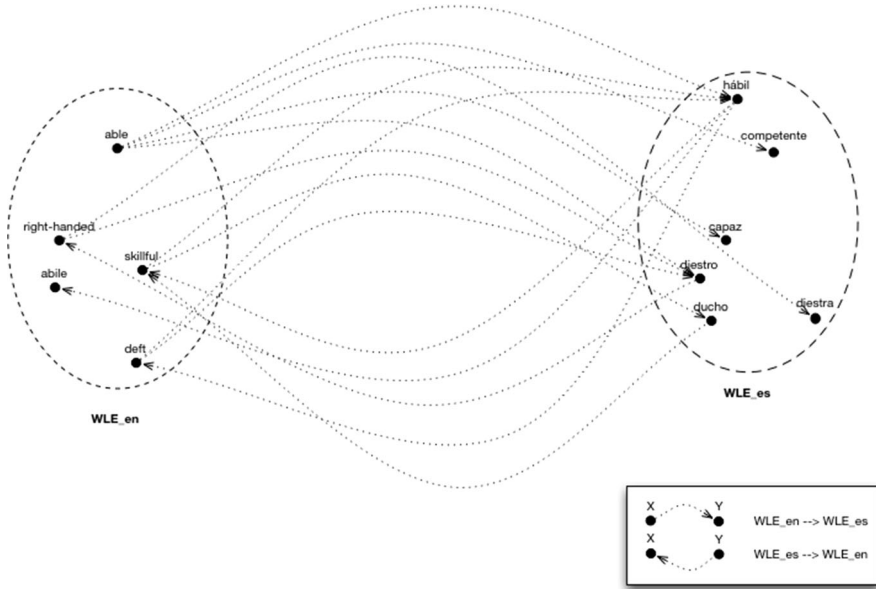


Fig. 3 $ISG_{en,es}(able)$

In the same way, we define:

$$arc_{source \rightarrow target}(w_i^{source}) = \{e_{i,1}^{source \rightarrow target}, \dots, e_{i,m}^{source \rightarrow target}\}$$

as the function returning the set of edges for a given word in WLE_{source} with inter-wiki links to WLE_{target} . Note that this set represents in fact the set of inter-wiki links.

Let $V_i^{source,target}$ and $E_i^{source,target}$ be the set of all words and links after applying f and arc to the word w_i^{source} and to each word in $\{w_1^{target}, \dots, w_m^{target}\}$ recursively, we define:

$$ISG_{source,target}(w_i^{source}) = \{V_i^{source,target}, E_i^{source,target}\}$$

w_i^{source} is known as the seed word for generating the ISG.

In order to build the network that is going to be used as input to the PageRank, we merge all ISGs under a Unified Semantic Graph (USG):

$$USG_{source,target} = \bigcup_{1 \leq i \leq n} \{ISG_{source,target}(w_i^{source})\}$$

We define a translation pair containing a word w_i^{source} from WLE_{source} and a word w_j^{target} from WLE_{target} as:

$$t_{i,j}^{source,target} = (w_i^{source}, w_j^{target})$$

Note that $(w_i^{source}, w_j^{target}) \equiv (w_j^{target}, w_i^{source})$, since “translation of” is a symmetric relationship.

We define the set of all translation pairs in WLE_{source} and WLE_{target} as:

$$T_{source,target} = \left\{ t_{i,j}^{source,target} \mid 1 \leq i \leq n; 1 \leq j \leq m \right\}$$

T does not contain repeated translations, i.e., in the case of symmetrical pairs we only consider one translation. Our proposed algorithm assigns a confidence value in the range [0–1] to each translation pair within $T_{source,target}$. For doing, so we apply the PageRank algorithm to the nodes of the $USG_{source,target}$ graph. This results in the association of a PageRank score to every node of the graph, i.e., w_i^{source} or w_j^{target} , independently. As we are interested about a confidence score for the translation pairs, we need to combine both PageRank values by using the following expression:

$$Score\left(t_{i,j}^{source,target}\right) = \frac{PR(w_i^{source}) + PR(w_j^{target})}{2}$$

This expression computes the mean of the PageRank values associated to each component of the pair. This value represents the confidence score associated to each translation pair by our algorithm.

3.2 Evaluation

In order to evaluate our approach we applied it to a subset of Wiktionary. We used the Ogden’s basic English word list⁸ consisting on 850 selected words. We used this list as a seed for generating the USG containing the translation pairs to rank and get the confidence values.

The aim of choosing a predefined list of words is to have a reference in order to compare with future approaches. Nevertheless, our algorithm can be adapted to take as input a word list generated from a chosen Wiktionary dump containing the latest available data. Our experiment relies uniquely on English and Spanish, but it can be easily extended to other language combinations as well.

3.2.1 Accuracy estimation

Given the Ogden’s list of seed words, we build the $USG_{en,es}$ and compute the confidence values for each translation pair as described previously. Figure 4 shows the histogram of the confidence values, which contains a total of 7366 nodes. The resulting dataset has approximately 12k translation pairs. As it can be appreciated, the most frequent values are placed near the mean (0.0001357).

⁸ <http://ogden.basic-english.org>.

We then label each pair using the following binary classification:

$$g(t) = \begin{cases} 1, & \text{if } \text{score}(t) \geq \frac{1}{\#\text{nodes in USG}} \\ 0, & \text{otherwise} \end{cases}$$

In this expression, the value 1 represents a translation pair that is commonly accepted. The value 0, on the other hand, represents a translation that is rare. Note that we use the value $1/(\#\text{nodes in USG})$ as the classification factor, because it is the average value computed by PageRank. Recall that the computed PageRank vector represents an N-dimensional probability distribution, which means that the total of all the probabilities must add up to 1. We use this binary classification with the aim of measuring the accuracy of our approach towards judgments generated manually.

We selected a total of 10 people, 5 native Spanish speakers and 5 native English speakers, with an intermediate or superior level in the respective foreign language (a minimum B2 level of the CERF⁹ was required). We prepared a survey consisting on 100 translations pairs extracted randomly from the total of approximately 12k generated pairs. We asked the volunteers to label each translation pair with either *common* or *rare* according to their knowledge. The surveys were made available online via Google Forms and anonymously submitted. In order to consider a survey as valid, all the 100 responses must be filled in. With all received submissions we calculated the average response vector and used this one for our comparison.

Additionally, we used a random baseline in our comparison simulating 500 independent evaluators. Each evaluator is represented by a vector of 100 elements, being each element either a 0 or a 1. We computed the average of the 500 vectors to extract the final vector used for labeling the translation pairs. Table 1 shows the 100 translation pairs together with the computed confidence. In addition, each pair shows the label generated by following each of the previous approaches.

Table 2 shows the precision and recall for our approach and the random baseline. While the precision of our system is higher than the obtained with the random approach, the recall is nearly the same. The reason for the low recall can be found on the way we compute the confidence for the pair by combining the individual PageRank values of the words. As stated previously, we obtained the final score by computing the average of both values. This assumes that both words have the same weight in the final confidence score. We observed cases like the pair (*hueso*, *bone*), with individual PageRank values of (8.580229958181926e-05, 0.00011250231809933178), which shows that *bone* is a lot more relevant than *hueso*. This means that the structure of one language edition can increase or penalize the final confidence value. A future line of improvement would consist on evaluating if the structure of the ISGs could be used to extract weights that help tuning the final confidence scores.

In any case, the use of a binary classification was for the sake of measuring the accuracy of our system. The main purpose of the computed scores is to be used as a weighted score. That is, instead of using each translation score as a binary value (common vs. rare), use it as a gradual measure to get an estimation of the confidence.

⁹ http://www.coe.int/t/dg4/linguistic/Cadre1_en.asp.

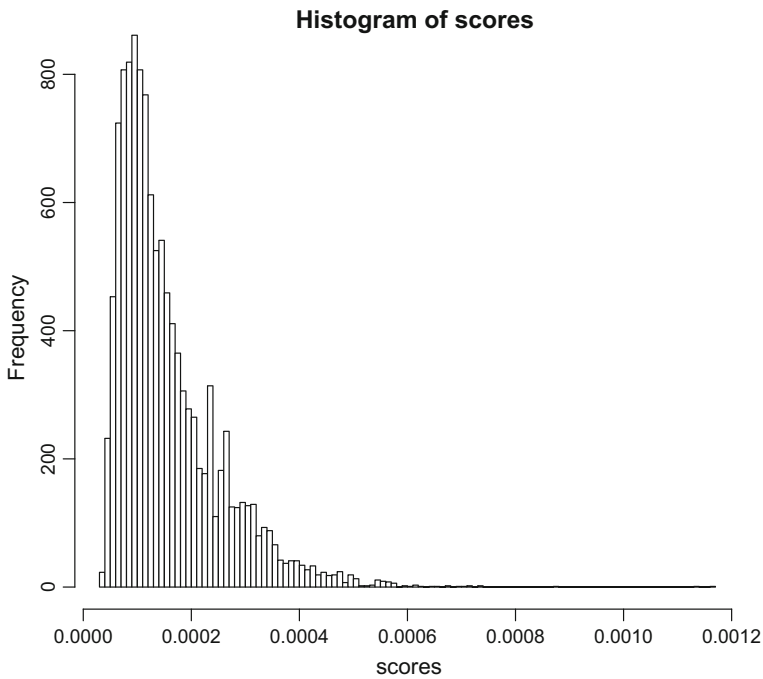


Fig. 4 Histogram for the confidence values extracted from *USG_{en,es}*

4 Sharing translations as Linked Open Data

Enabling interoperability in a LOD environment of different NLP systems that perform complementary tasks would facilitate the comparison of results, as well as the combination of tools to build more complex and reusable processes. A step towards NLP interoperability relies on the definition of shared vocabularies to handle linguistics. In this section we focus on the definition of a common vocabulary for modeling linguistic data, with special attention to multilingual translations.

At the time of writing, *lemon*¹⁰ (McCrae et al. 2012) can be considered as de facto vocabulary for modeling linguistics on the LOD cloud. In order to avoid reinventing the wheel by proposing a new model, we will reuse *lemon* as much as possible and try to extend those components that we might need to fulfill our requirements.

4.1 Overview of the *lemon* model

The LEXical Model for ONtologies (*lemon*) is an RDF model to describe linguistic information associated to ontologies. One of the main features of *lemon* is that it

¹⁰ <http://www.lemon-model.net>.

Table 1 Evaluation results

	w_1	w_2	Confidence	Link-based label	Random label	Expert label
1	letra	lyrics	0.00015808254846777422	1	1	1
2	alejar	remove	0.00011903837408350023	0	1	0
3	homosexual	homosexual	0.00028793665648534366	1	1	1
4	duro	hard	0.00011151357239946473	0	0	1
5	replace	reemplazar	0.000272233906003052	1	1	1
6	university	universidad	0.0003312782931114376	1	1	1
7	loose	golfo	0.00013784968353348345	1	0	0
8	grand	grande	0.00011685515559887251	0	1	0
9	back	anterior	0.00022811473566881207	1	0	1
10	vivienda	house	0.0001259226010898903	0	1	0
11	kiss	beso	0.0002681398147512756	1	0	1
12	nobody	pelagatos	0.0001872058505674174	1	1	1
13	bajo	under	0.0003398497804401553	1	1	1
14	speak	hablar	8.565929788258152e-05	0	1	1
15	hammer	percutor	9.392650280119472e-05	0	1	0
16	cavalry	caballera	0.00027503509647252655	1	0	1
17	accessory	accesorio	4.748358294852667e-05	0	1	1
18	arreglar	arrange	0.00015134654222971389	1	0	1
19	book	libreto	5.542939908986271e-05	0	0	0
20	encoder	light	0.0001292829908069997	0	1	1
21	shit	puta	0.00016815678522440745	1	1	0
22	dude	parce	9.605754355215815e-05	0	1	0
23	deed	acta	7.735300293145157e-05	0	0	1

Table 1 continued

	w_1	w_2	Confidence	Link-based label	Random label	Expert label
24	cover	ocular	0.00019365656483481532	1	1	1
25	glue	pegamento	9.439440781893953e-05	0	0	1
26	biscuit	bizcocho	0.0003315274781177747	1	1	1
27	francés	french	8.131212808787852e-05	0	0	1
28	drink	trago	0.00017873326871777366	1	1	1
29	lift	subir	0.00017888264071815794	1	1	1
30	straighten	desencorvar	5.589172980399047e-05	0	0	0
31	how	cómo	0.00012595489284186208	0	0	1
32	wander	engañar	0.0001165285131188344	0	1	0
33	morning	mañana	0.0001020673285550087	0	0	1
34	family	familiar	0.0004261312487014779	1	1	1
35	married	casado	0.00023955231483800904	1	1	1
36	search	búsqueda	0.0003157591556240658	1	0	1
37	in love	enamorado	0.00011433544834685391	0	0	1
38	tail	cola	0.00012383394442699207	0	0	1
39	transacción	transaccion	0.0002491944717668268	1	0	1
40	all right	listo	0.00013938872246804576	1	1	0
41	suerte	chance	0.00018970906196004133	1	0	0
42	error	error	0.0001828805257883463	1	0	1
43	enfadado	angry	0.00020231485212181724	1	1	1
44	shrub	arbusto	0.0003373185713546105	1	0	1
45	bank	ingresar	7.283068865804214e-05	0	1	0
46	bud	brotar	5.2801137963791824e-05	0	1	1
47	screwed	jodido	9.776830619674182e-05	0	1	1

Table 1 continued

	w_1	w_2	Confidence	Link-based label	Random label	Expert label
48	claro	bright	6.99520357867713e-05	0	1	1
49	ver	see	0.00030771094402188724	1	0	1
50	mermelada	fruit preserve	7.375930078031526e-05	0	1	0
51	fijar	fix	0.00013327324024832796	0	0	1
52	cayo	key	0.00020364632506934789	1	0	0
53	dormir	sleep	0.00013320207076399307	0	1	1
54	flow	caudal	0.00012077135530924626	0	1	1
55	education	educación	0.00023955231483800904	1	1	1
56	cuerda	cord	9.187920825884666e-05	0	0	1
57	political	político	5.204163558498278e-05	0	1	1
58	alambique	still	8.803964940052668e-05	0	0	0
59	spinning top	peonza	8.753165645304814e-05	0	1	1
60	step	peldaño	0.00015230951893320462	1	1	1
61	pit	excavación	0.00012815550970501842	0	0	1
62	throw	tirar	0.0002091710041445345	1	0	1
63	drop	gota	0.0001661758296227969	1	0	1
64	frame	encuadrar	7.662048838106785e-05	0	1	1
65	señorita	Miss	0.00014043014399257818	1	0	1
66	tart	ácido	0.0002006871967378895	1	1	1
67	furnace	caldera	0.00010816230756751657	0	0	1
68	judge	opinar	0.0002118889527781566	1	0	1
69	pelado	nudo	6.824036805736078e-05	0	0	0
70	cause	causar	0.0004612281806414062	1	1	1

Table 1 continued

	w_1	w_2	Confidence	Link-based label	Random label	Expert label
71	junk	cachivache	7.25134795117984e-05	0	1	1
72	unidad	unity	0.0002681398147512756	1	1	1
73	floor	piso	9.04091833947343e-05	0	1	1
74	crop	buche	6.51696734496451e-05	0	1	0
75	want	desear	0.00027765843693083436	1	0	1
76	filthy	mugriento	6.40411203511864e-05	0	0	1
77	hardheaded	cabezadura	7.617219715207323e-05	0	0	1
78	sweet	rico	0.00012041507524361786	0	1	0
79	organization	organización	0.00025605162014417015	1	1	1
80	start	comienzo	0.00022775054735326602	1	1	1
81	insinuar	hint	0.00015056975574942656	1	1	1
82	wealth	riqueza	0.00011482456884216536	0	0	1
83	lila	lilac	0.0003320244735367959	1	0	1
84	mister	seor	0.00011594181554690203	0	1	1
85	mass	música	0.00013804712258276455	1	1	0
86	wash	enjuague	9.82630154478201e-05	0	0	1
87	the	las	0.0001877848958416193	1	1	1
88	clean	limpiar	0.00019142997506303304	1	0	0
89	hueso	bone	9.915230884057551e-05	0	1	1
90	drink	tomar	0.0002229550318493597	1	0	1
91	back	lomo	0.00025485828009033873	1	0	1
92	war	guerrear	9.392650280119472e-05	0	1	1
93	claim	sostener	0.0001855549283132706	1	0	1
94	keep	seguir	0.0002610192248416121	1	1	1

Table 1 continued

	w_1	w_2	Confidence	Link-based label	Random label	Expert label
95	bundle	haz	9.788103215336511e-05	0	0	1
96	sentir	regret	8.114451285740591e-05	0	0	0
97	amar	love	5.778333304465147e-05	0	0	1
98	creencia	belief	9.392650280119472e-05	0	0	1
99	plane	avión	0.00017110862695491545	1	0	1
100	loco	insane	7.233687654911745e-05	0	1	1

keeps the independence between the target ontology and the linguistic descriptions. This means that the use of an ontology is not compulsory in order to create linguistic descriptions. This design factor is pushing the *lemon* vocabulary to become de facto model for describing and exchanging linguistic resources on the Web.¹¹

The different modules¹² that compose the *lemon* vocabulary are the following:

- *Core* as the name indicates, this module is the central part of the lemon model. It provides the required constructions to represent lexical descriptions and associate them to the concept of an external ontology. Figure 5 shows a representation of the vocabulary constructions available in this module. The class *Lexicon* is used to encapsulate lexical descriptions referring to certain preferences. For example, lexical entries belonging to the English language. In this way, *lemon* allows the definition of sets of descriptions grouped independently.
- *Linguistic description* this module allows the usage of properties for adding information to the lexicon like part-of-speech, gender, number, tenses, phonetics, etc.
- *Variation* this module provides the necessary vocabulary constructions to build relationships among the element of a lexicon, for example synonyms, antonyms and translations.
- *Phrase structure* several constructions are provided in this module in order to deal with multiple word expressions.
- *Syntax and mapping* this module contains the needed vocabulary to establish syntactic rules between lexical components.
- *Morphology* the target of this module is to handle the different forms of a lexical entry.

4.2 Coping with translations

Several approaches have been developed to annotate ontologies with natural language descriptions, such as the *rdfs:label* (Manola and Miller 2004) or *skos:prefLabel* (Miles and Bechhofer 2009b) properties. These properties allow the use of simple multilingual labels (e.g., *rdfs:label* “*mountain*”@en). However, the main limitation of this approach is the impossibility to create explicit links among labels. In order to solve this limitation SKOS-XL (Miles and Bechhofer 2009a) introduced a *skosxl:Label* class that allows labels to be handled as RDF resources. In addition, a *skosxl:labelRelation* property was introduced to create links between instances of *skosxl:Label*, which could be used for establishing translation relationships.

¹¹ The latest developments on *lemon* by part of the Ontology Lexicon (Ontolex) community group can be found at <http://www.w3.org/2016/05/ontolex/>.

¹² For a detailed description of the different *lemon* components we refer the reader to the official cookbook at <http://www.lemon-model.net/lemon-cookbook>.

Table 2 Precision-recall comparison for our approach and a random baseline

	Link analysis	Random
Precision	0.8775	0.7592
Recall	0.5512	0.5256

The lemon core model allows the implicit representation of multilingual information, in the sense that several lexicon models using different natural languages can reference to the same ontology. We say it is an implicit representation because translation relations can only be inferred when they point to the same ontology entity. In case that more information about translations must be stated or when the references to the ontology from the lexical entries are not available, additional approaches must be considered in order to explicitly represent translations as RDF resources. The advantage of introducing explicit translations is that the model is more independent of the target ontology.

lemon does not contemplate the explicit handling of translations in its core model. Flourishing efforts try to keep independent of any ontology and therefore implement an explicit model for the translations. A first step in this direction has been implemented in the lemon variant module, which introduces a *senseRelation* property that can be used to create relationships between senses. As a specialization of this property, the *isocat:translationOf* property is used to create translation links between different senses of words. The translation link itself does not include information about the language pairs that compose the translation. The only way to extract the languages is by using the lexicons containing the terms.

Montiel-Ponsoda et al. (2011) introduce an extension module to *lemon* for modeling translations explicitly. The main part of this module is the *Translation* class, which represent the relation between lexical senses. Instances of the *lemon:LexicalSense* class composing the translation pair are referenced by two new properties, namely *sourceLexicalSense* and *targetLexicalSense*. Authors differentiate two kinds of translations, i.e., *LiteralTranslation* and *CulturalEquivalenceTranslation*. The first type is used in the case the semantic of the sense is equivalent in both languages, e.g.: “mountain”@en and “Berg”@de. The second type of translation is used when the lexical senses can not be considered exactly equivalent, but only under certain cultural context, e.g.: “Prime Minister” and “Presidente del Gobierno” in the British and Spanish political systems, respectively. Figure 6 shows an overview of this model.

Gracia et al. (2014) propose a modification¹³ of the work described in Montiel-Ponsoda et al. (2011). Figure 7 shows an overview of the model. The main difference relies on the introduction of a *TranslationSet* class, which groups the different translations in a similar way to how *lemon:Lexicon* groups lexical entries. The *Translation* class has been modified, so that the original specialization classes, i.e., *LiteralTranslation* and *CulturalEquivalenceTranslation*, have been removed. However, that information can still be modeled by using a *translationCategory*

¹³ At the time of writing, this modification is being considered as a possible approach to model explicit translations in lemon. Further discussions are available at http://www.w3.org/community/ontolex/wiki/Translation_Module.

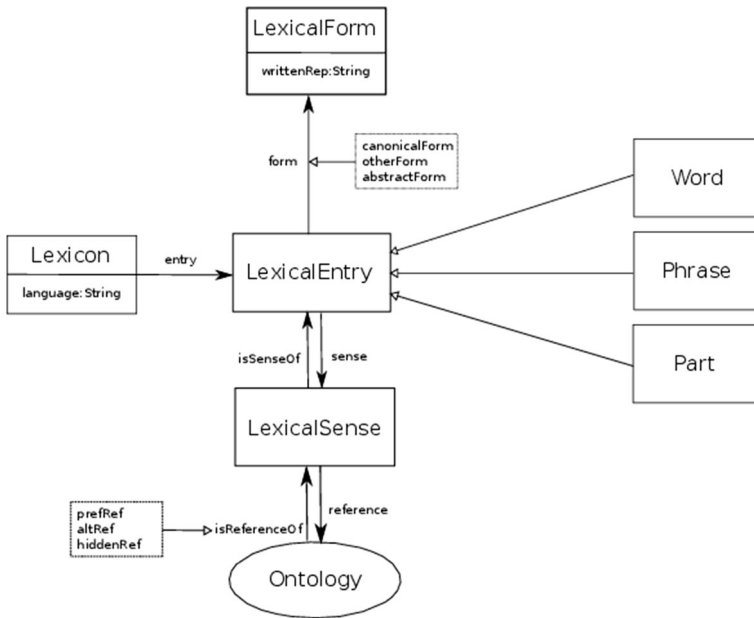


Fig. 5 Lemon core model. (source: <http://www.lemon-model.net>)

property that points to an external registry of translation types.¹⁴ At the time of writing, authors suggest the following categories:

- *directEquivalent* Typically, the two terms describe semantically equivalent entities that refer to entities that exist in both cultures and languages. E.g., “surrogate mother”@en and “mère porteuse”@fr.
- *culturalEquivalent* Typically, the two terms describe entities that are not semantically but pragmatically equivalent, since they describe similar situations in different cultures and languages. E.g., “Ecole Normal”@fr and “Teachers college”@en.
- *lexicalEquivalent* It is said of those terms in different languages that usually point to the same entity, but one of them verbalizes the original term by using target language words. E.g., “Ecole Normal”@fr and “Normal School”@en.

In addition a *context* property has been added to *Translation*, which points to extra information that specifies the concrete context in which the pair of lexical senses compose the translation. This information can be really important when disambiguating senses of a word.

Sérasset (2014) introduce DBnary, a multilingual LOD dataset¹⁵ built using Wiktionary as data source. DBnary relies on *lemon* and adds an extension for

¹⁴ The category registry can be found at <http://purl.org/net/translation>.

¹⁵ The dataset is available at <http://kaiko.getalp.org/sparql>.

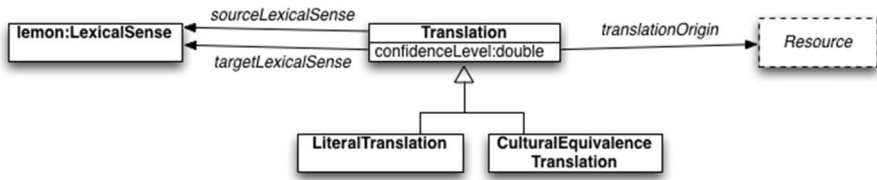


Fig. 6 Translation handling model suggested in Montiel-Ponsoda et al. (2011)

dealing with the information in every Wiktionary page. From all lexical relations that might appear in Wiktionary and are handled in DBnary, we only focus on translations. To make translations explicit a *Translation* class is defined. This class provides the following set of properties:

- *isTranslationOf* Relates a *Translation* to a *LexicalEntry* or *LexicalSense*. It is important to remark the difference with the model in Montiel-Ponsoda et al. (2011) and Gracia et al. (2014), where a *sourceLexicalSense* and a *targetLexicalSense* properties are used to describe the translation pair. In DBnary only a relation is needed to indicate the source of the translation because the target is modeled by using the string property *writtenForm*. This fact implies redundancy of information in DBnary and will require the consolidation between the information pointed by *writtenForm* and the one included in *LexicalEntry* or *LexicalSense*, respectively.
- *targetLanguage* Points to one of the ISO 639-3 language codes available in the *lexvo*¹⁶ namespace.
- *writtenForm* Gives the string representation of the translation in the target language.
- *gloss* Is a string containing information that determines the context of the lexical sense. It is similar to the *context* property introduced in Gracia et al. (2014).
- *usage* It is a string that gives extra information about the translation, i.e., genre, number, etc.

Figure 8 depicts an overview of the model.

4.3 Modeling confidence and provenance

Montiel-Ponsoda et al. (2011) and Gracia et al. (2014) remark the need for modeling provenance and confidence scores associated to translations. The reason for this is the ambiguity of lexical senses (usually a word in a certain language could have different senses, which could be translated differently into a target language) and the implicit subjectivity that characterizes the translation of lexical content, i.e., lack of accuracy of automatic methods, lack of consensus among human contributors, etc.

The model in Montiel-Ponsoda et al. (2011) allows the specification of provenance information associated to translations by using the *translationOrigin*

¹⁶ <http://www.lexvo.org>.

property for pointing to external resources. Gracia et al. (2014) provenance is also taken into account through the use of DCMI metadata terms,¹⁷ associated to *Translation* and *TranslationSet* instances

Additionally, in Montiel-Ponsoda et al. (2011) confidence values can be added to a translation by using the *confidenceLevel* property [renamed to *translationConfidence* in Gracia et al. (2014)]. Confidence can be interpreted as a ranking score denoting how trustable the defined translation is. Authors state that the “confidence level will ultimately depend on the translation tools and translation resources employed to obtain translations”. However, they do not include any kind of vocabulary construction in their approach to model information about the applied tool. Ranking scores themselves do not provide enough information that can be used during the data consumption process. With this idea in mind, we developed a vocabulary for ranking (*vRank*), which allows the reification of ranking data. We will use *vRank* for modeling confidence associated to lexical translations. *vRank* has been designed in order to facilitate its reusability and therefore it is independent of the targeted use case.

The purpose of *vRank* is to provide data consumers with a standardized, formal, unambiguous, reusable and extensible way of representing ranking computations. How data is consumed depends strongly on what is relevant for data consumers. When data providers offer a ranking service, obviously they cannot contemplate all possible relevance models of consumers. Therefore, the need for the functionality that *vRank* tries to implement is apparent, i.e., associating different ranking models with the same data. The following requirements have guided the design of *vRank*:

1. The need for unifying the way ranking algorithms are developed in order to promote reusability and evaluation.
2. The need for a common and accepted model to homogenize the exploitation of ranking services.
3. The need for isolating data from any kind of assumption regarding to publication and consumption (data providers and consumers may not share the same interests).

Offering ranking computations as part of the data can facilitate its consumption in several ways:

- Different relevance models computed by diverse ranking strategies can coexist within the same dataset. Consumers can adapt data requests to their relevance expectations.
- Data ranking becomes open and shareable. Consumers can reuse a specific way of ranking a dataset. If existent ranking approaches do not suit consumers' needs, they can extend the dataset with their own method.
- Consumers can reuse ranking scores in order to evaluate and compare different strategies over a given dataset.

¹⁷ <http://dublincore.org/documents/dcmi-terms/>.

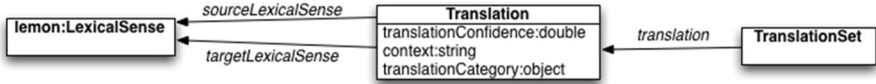


Fig. 7 Translation handling model suggested in Gracia et al. (2014)

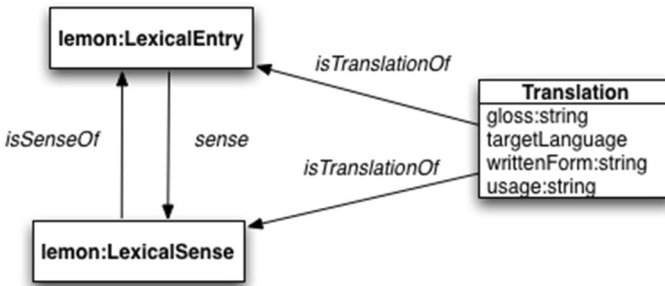


Fig. 8 Translation handling model suggested in Sérasset (2014)

- Consumers (and not data providers) can have control about how they want to consume data, giving more preference to what is more relevant.

In the following we develop these aspects in more detail.

Ranking crystallization Ranking algorithms rely on data structures that are used to compute the final scores of data items. Traditionally, these data structures are kept internal and inaccessible for the users. By using a service, data consumers can submit their queries and retrieve a list of results ordered according to the implemented relevance model. This kind of behavior defines ranking algorithms as a black box, which makes it very difficult, if not impossible, to reuse and share computations over existing data.

The relevance models computed by ranking algorithms need to be materialized in a way that can be offered publicly and can be queried by data consumers. The publication can be done “easily” in RDF with a vocabulary that models the ranking domain. This is what *vRank* has been defined for.

SPARQL¹⁸ is considered as the standard language for querying the Web of Data, however it does not support any kind of ranking apart from ORDER BY clauses. By adopting *vRank* it is not necessary to extend SPARQL with ranking support as the ranking can be made explicit within the dataset. Consumers do not need to learn a different query language or any kind of extension. They still can use ORDER BY clauses and just adapt their queries to use the according *vRank* triples.

Ranking evaluation Due to the different policies used in ranking, it is very difficult to establish a technical comparison to analyze the accuracy and precision of each algorithm in reference to others. One of the main contributions of *vRank* is that

¹⁸ <http://www.w3.org/TR/rdf-sparql-query/>.

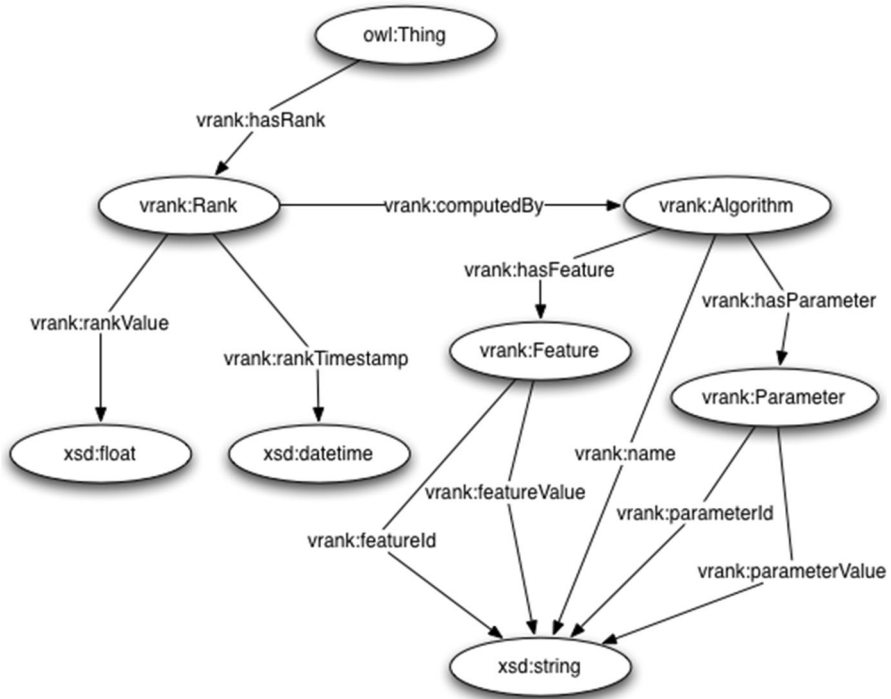


Fig. 9 vRank overview

it helps to homogenize the way ranking services are exploited, so that third parties can compare and evaluate them.

Evolution of data In an open environment like the Web, data is always going under modifications and revisions. When data is updated, the ranking scores associated to the data items have to be updated as well. A consumer may be interested in analyzing the ranking scores over the time in order to predict future changes that might affect her consumption patterns. The mechanism implemented by *vRank* opens new possibilities for addressing the problem of measuring changes within data.

Multirelevancy Consumers can make use of the available ranking scores to combine and compose their own ranking functions. This approach is addressed under what is known in the literature as ranking aggregation (Dwork et al. 2001). Following the same pattern, the newly obtained scores can be materialized and shared by using *vRank*.

4.3.1 *vRank* model

vRank aims to model ranking information within datasets. We have tried to keep a simplistic design and therefore, we have reused existent vocabularies wherever possible. A full specification of *vRank* is available under the namespace <http://>

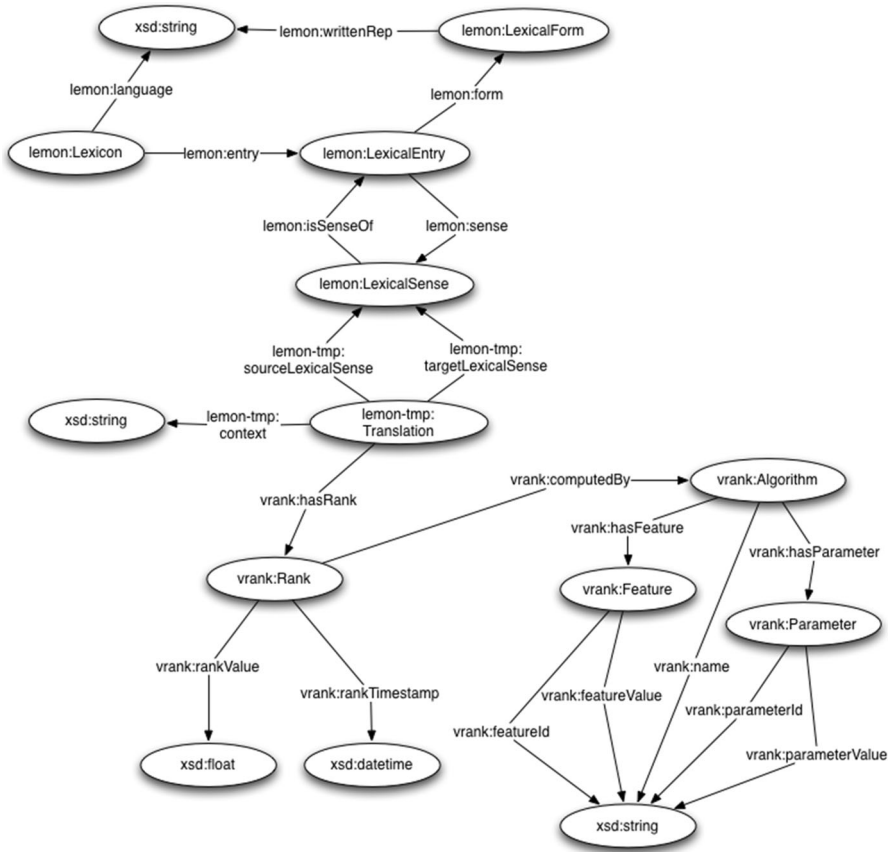


Fig. 10 Overview of the proposed lemon extension

purl.org/voc/vrank#. Figure 9 shows an overview of *vRank*. In the following, we describe the core components of the vocabulary.

- *Algorithm*: In *vRank* an *Algorithm* is an entity that models metadata about a ranking implementation. The main purpose of this entity is to provide provenance information about the ranking scores. By knowing which settings have produced certain ranking scores, a data consumer can decide which ranking approach should be applied to the requested data. In order to characterize certain algorithm, *vRank* allows the use of features and parameters.
- *Feature*: A *Feature* complements the description of an algorithm in terms of its functionality. Features should be specified by the authors of the ranking approach with the aim of facilitating its understanding to data consumers. As already mentioned, ranking algorithms are characterized by a diverse functionality, which in many cases is combined under the same implementation.

- *Parameter*: A *Parameter* adds a finer level of description than a *Feature*. The main target of a *Parameter* is to capture the specific configuration of the algorithm that leads to the obtained ranking scores. An example of *Parameter* is the damping factor used by PageRank.
- *Rank*: *Rank* is an entity that formalizes the ranking scores associated to a data item. Anything that can be model in RDF can have an associated *Rank*. The flexibility of the model resides on relating different instances of *Rank* with a particular data item. A *Rank* by itself is meaningless. Therefore, *Ranks* are related to *Algorithms* and to concrete executions (defined by specifying different *Parameters*). In order to capture different executions with certain settings we have added a timestamp to the *Rank* entity.

4.3.2 Consolidated model

As already stated, our aim is to reuse accepted models by the community in order to avoid reinventing the wheel with a new approach. Figure 10 shows a representation of the resulting vocabulary that we use for modeling the multilingual translations with the associated confidence annotations. We have used the prefix *lemon-tmp* to refer to the vocabulary that is related to the translations. As it can be appreciated, we have relied on previous works addressing the description of translations as LOD, especially (Gracia et al. 2014). A translation is modeled through the entity *lemon-tmp:Translation*, which contains two relations to *lemon:LexicalSense* for representing the source and the target senses of the translation pair. This information is modeled through the properties *lemon-tmp:sourceLexicalSense* and *lemon-tmp:targetLexicalSense*, respectively. The confidence information related to each translation pair is modeled through the property *vrank:hasRank*, which serves as junction between *lemon* and *vRank*.

4.4 Resulting dataset

We have used the model introduced previously to represent all the generated information about translations and their confidence after running the evaluation described in Sect. 3.2. A generic overview of the dataset is shown in Fig. 11. For each couple of processed language editions we organize the data in a *source lexicon*, a *target lexicon* and a *translation set*. The *source lexicon* and *target lexicon* contain information about words in the source and target language, respectively. In the current version of the dataset we just include the lexical form of each word, discarding other lexical information like definitions, synonyms, etc. The *translation set* includes all the translation pairs, together with the associated confidence score computed with our approach. We made this English–Spanish dataset containing more than 100k triples publicly available on Dydra¹⁹

¹⁹ <http://dydra.com>.

under the following SPARQL endpoint: <http://dydra.com/narko/dict/sparql>. An example of the dataset is described in the “Appendix”.

5 Conclusions and future work

This paper pointed out the need for establishing measures to show the level of reliability of linguistic open data. More concretely, we studied the multilingual dimension of Wiktionary in order to assess the quality of translations as a previous step before using them in practice. We proposed a heuristic approach based on random walks to exploit the link structure existing between different Wiktionary language editions. This mechanism yielded confidence values associated to each translation pair found within the network formed by combining a source and a target language edition. We studied the precision and recall of our approach evaluating towards human assessments and found out that the heuristic performs significantly good.

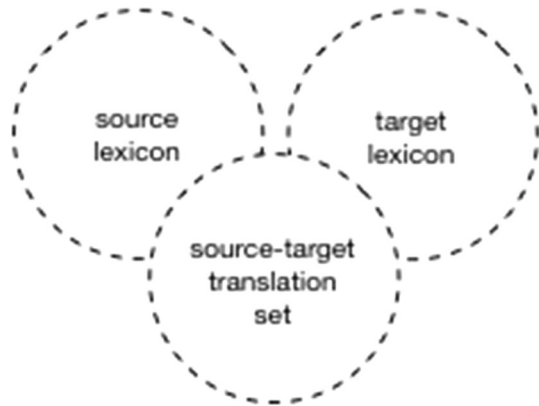
As a complementary part of this paper, we extended the state of the art by providing an extension to the *lemon* vocabulary to model translation data with associated confidence measures. We showed the flexibility of LOD to model new data needs and how it can overcome heterogeneity issues that are present in fixed data schemas like those used in Wiktionary. As a proof of concept we published a dataset containing all translation pairs we computed by relying on the described vocabulary.

In the short term, future work includes the improvement of the weighting mechanism used to compute the confidence measures associated to the translation pairs. As described, right now we rely on the mean of the individual translation components. However, this has shown to have a high impact in the case the associated ISG to one of the components of the pair is incomplete. This effect appears in those words that did not receive enough contributions within a language edition. Our initial idea is to suggest weights that penalize the component of the pair suffering this problem, while increasing the relevance of the other component. Additionally, we want to apply our approach to any combination of Wiktionary language editions in order to provide the community with one of the biggest datasets containing bilingual translations. This resource can be quite useful for researchers on applied linguistics and semantics.

In the long term, we consider applying our heuristic to other linguistic resources apart of Wiktionary. Projects like UBY²⁰ and BabelNet²¹ integrate other multilingual resources that could be explored. These projects could make use of our approach in order to offer confidence values associated to the translations that they store. As already stated through the paper, our approach relies strongly on the link structure built among the different Wiktionary editions. Applying our heuristic to a different resource in order to capture the crowd contribution would require the implementation of a different data analysis targeting such resource. This means that

²⁰ <https://www.ukp.tu-darmstadt.de/data/lexical-resources/uby/>.

²¹ <http://babelnet.org/>.

Fig. 11 Conceptual overview of the generated dataset**Table 3** Adjacency matrix corresponding to $ISG_{en,es}(able)$ and associated confidence

Source	Target	Score
deft	diestro	0.099882
deft	hábil	0.102131
skillful	ducho	0.125119
skillful	hábil	0.143478
diestro	right-handed	0.135940
able	competente	0.035715
able	hábil	0.083146
able	capaz	0.035715
able	diestro	0.080898
right-handed	diestro	0.135940
right-handed	diestra	0.117580
ducho	skillful	0.125119
hábil	skillful	0.143478
hábil	abile	0.102131
hábil	deft	0.102131

while the main idea behind our heuristic can be reused, the current technical implementation would need to be adapted to each targeted resource.

Appendix: Dataset example

In the following, we show an example of how our data model can be used to describe lexical translations. We have taken the word “able” in English and build the associated ISG for Spanish as described in Sect. 3. The resulting graph is shown in Fig. 3. Table 3 contains the adjacency matrix with the existing translations and the computed confidence after combining the individual PageRank scores. Note that for this example we take the ISG as the only graph under consideration and therefore it is equivalent to the USG. Listing 5 depicts the generated model in turtle notation.

```

@prefix lemon: <http://lemon-model.net/lemon#> .
@prefix lemon-tmp: <http://purl.org/voc/lemon-tmp#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix vrank: <http://purl.org/voc/vrank#> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

lemon:lexicalEntry_10__capaz a lemon:LexicalEntry ;
    lemon:form lemon:lexicalForm_10 ;
    lemon:sense lemon:lexicalSense_10__capaz .

lemon:lexicalEntry_11__abile a lemon:LexicalEntry ;
    lemon:form lemon:lexicalForm_11 ;
    lemon:sense lemon:lexicalSense_11__abile .

lemon:lexicalEntry_1__deft a lemon:LexicalEntry ;
    lemon:form lemon:lexicalForm_1 ;
    lemon:sense lemon:lexicalSense_1__deft .

lemon:lexicalEntry_2__diestro a lemon:LexicalEntry ;
    lemon:form lemon:lexicalForm_2 ;
    lemon:sense lemon:lexicalSense_2__diestro .

lemon:lexicalEntry_3__habil a lemon:LexicalEntry ;
    lemon:form lemon:lexicalForm_3 ;
    lemon:sense lemon:lexicalSense_3__habil .

lemon:lexicalEntry_4__skillful a lemon:LexicalEntry ;
    lemon:form lemon:lexicalForm_4 ;
    lemon:sense lemon:lexicalSense_4__skillful .

lemon:lexicalEntry_5__ducho a lemon:LexicalEntry ;
    lemon:form lemon:lexicalForm_5 ;
    lemon:sense lemon:lexicalSense_5__ducho .

lemon:lexicalEntry_6__right-handed a lemon:LexicalEntry ;
    lemon:form lemon:lexicalForm_6 ;
    lemon:sense lemon:lexicalSense_6__right-handed .

lemon:lexicalEntry_7__diestra a lemon:LexicalEntry ;
    lemon:form lemon:lexicalForm_7 ;
    lemon:sense lemon:lexicalSense_7__diestra .

lemon:lexicalEntry_8__able a lemon:LexicalEntry ;
    lemon:form lemon:lexicalForm_8 ;
    lemon:sense lemon:lexicalSense_8__able .

lemon:lexicalEntry_9__competente a lemon:LexicalEntry ;
    lemon:form lemon:lexicalForm_9 ;
    lemon:sense lemon:lexicalSense_9__competente .

lemon:lexicalSense_10__capaz a lemon:LexicalSense .
lemon:lexicalSense_11__abile a lemon:LexicalSense .
lemon:lexicalSense_7__diestra a lemon:LexicalSense .
lemon:lexicalSense_9__competente a lemon:LexicalSense .
lemon:lexicalSense_5__ducho a lemon:LexicalSense .
lemon:lexicalSense_1__deft a lemon:LexicalSense .
lemon:lexicalSense_6__right-handed a lemon:LexicalSense .

```

```

lemon:lexicalSense_2__diestro a lemon:LexicalSense .
lemon:lexicalSense_4__skillful a lemon:LexicalSense .
lemon:lexicalSense_8__able a lemon:LexicalSense .
lemon:lexicalSense_3__habil a lemon:LexicalSense .
lemon:lexicalForm_1 lemon:writtenRep "deft"^^xsd:string .
lemon:lexicalForm_10 lemon:writtenRep "capaz"^^xsd:string .
lemon:lexicalForm_11 lemon:writtenRep "abile"^^xsd:string .
lemon:lexicalForm_2 lemon:writtenRep "diestro"^^xsd:string .
lemon:lexicalForm_3 lemon:writtenRep "habil"^^xsd:string .
lemon:lexicalForm_4 lemon:writtenRep "skillful"^^xsd:string .
lemon:lexicalForm_5 lemon:writtenRep "ducho"^^xsd:string .
lemon:lexicalForm_6 lemon:writtenRep "right-handed"^^xsd:string .
lemon:lexicalForm_7 lemon:writtenRep "diestra"^^xsd:string .
lemon:lexicalForm_8 lemon:writtenRep "able"^^xsd:string .
lemon:lexicalForm_9 lemon:writtenRep "competente"^^xsd:string .
lemon-tmp:translation_10__right-handed_diestro a lemon-tmp:Translation ;
  lemon-tmp:sourceLexicalSense lemon:lexicalSense_6__right-handed ;
  lemon-tmp:targetLexicalSense lemon:lexicalSense_2__diestro ;
  vranks:hasRank [ vranks:rankTimestamp "2014-09-01T10:22:57.303141"^^xsd:dateTime ;
    vranks:rankValue "0.135939716536"^^xsd:float ] .
lemon-tmp:translation_11__right-handed_diestra a lemon-tmp:Translation ;
  lemon-tmp:sourceLexicalSense lemon:lexicalSense_6__right-handed ;
  lemon-tmp:targetLexicalSense lemon:lexicalSense_7__diestra ;
  vranks:hasRank [ vranks:rankTimestamp "2014-09-01T10:22:57.303467"^^xsd:dateTime ;
    vranks:rankValue "0.117580422621"^^xsd:float ] .
lemon-tmp:translation_12__ducho_skillful a lemon-tmp:Translation ;
  lemon-tmp:sourceLexicalSense lemon:lexicalSense_5__ducho ;
  lemon-tmp:targetLexicalSense lemon:lexicalSense_4__skillful ;
  vranks:hasRank [ vranks:rankTimestamp "2014-09-01T10:22:57.304036"^^xsd:dateTime ;
    vranks:rankValue "0.125118537434"^^xsd:float ] .
lemon-tmp:translation_13__habil_skillful a lemon-tmp:Translation ;
  lemon-tmp:sourceLexicalSense lemon:lexicalSense_3__habil ;
  lemon-tmp:targetLexicalSense lemon:lexicalSense_4__skillful ;
  vranks:hasRank [ vranks:rankTimestamp "2014-09-01T10:22:57.304367"^^xsd:dateTime ;
    vranks:rankValue "0.143477831349"^^xsd:float ] .
lemon-tmp:translation_14__habil_abile a lemon-tmp:Translation ;
  lemon-tmp:sourceLexicalSense lemon:lexicalSense_3__habil ;
  lemon-tmp:targetLexicalSense lemon:lexicalSense_11__abile ;
  vranks:hasRank [ vranks:rankTimestamp "2014-09-01T10:22:57.304682"^^xsd:dateTime ;
    vranks:rankValue "0.102130513122"^^xsd:float ] .
lemon-tmp:translation_15__habil_deft a lemon-tmp:Translation ;
  lemon-tmp:sourceLexicalSense lemon:lexicalSense_3__habil ;
  lemon-tmp:targetLexicalSense lemon:lexicalSense_1__deft ;
  vranks:hasRank [ vranks:rankTimestamp "2014-09-01T10:22:57.304997"^^xsd:dateTime ;
    vranks:rankValue "0.102130513122"^^xsd:float ] .

```

```

lemon—tmp: translation_1__deft_diestro a lemon—tmp: Translation ;
  lemon—tmp: sourceLexicalSense lemon: lexicalSense_1__deft ;
  lemon—tmp: targetLexicalSense lemon: lexicalSense_2__diestro ;
  vrank: hasRank [ vrank: rankTimestamp "2014-09-01T10:22:57.297762"^^xsd:dateTime ;
    vrank: rankValue "0.0998822038435"^^xsd:float ] .

lemon—tmp: translation_2__deft_habil a lemon—tmp: Translation ;
  lemon—tmp: sourceLexicalSense lemon: lexicalSense_1__deft ;
  lemon—tmp: targetLexicalSense lemon: lexicalSense_3__habil ;
  vrank: hasRank [ vrank: rankTimestamp "2014-09-01T10:22:57.298734"^^xsd:dateTime ;
    vrank: rankValue "0.102130513122"^^xsd:float ] .

lemon—tmp: translation_3__skillful_ducho a lemon—tmp: Translation ;
  lemon—tmp: sourceLexicalSense lemon: lexicalSense_4__skillful ;
  lemon—tmp: targetLexicalSense lemon: lexicalSense_5__ducho ;
  vrank: hasRank [ vrank: rankTimestamp "2014-09-01T10:22:57.299575"^^xsd:dateTime ;
    vrank: rankValue "0.125118537434"^^xsd:float ] .

lemon—tmp: translation_4__skillful_habil a lemon—tmp: Translation ;
  lemon—tmp: sourceLexicalSense lemon: lexicalSense_4__skillful ;
  lemon—tmp: targetLexicalSense lemon: lexicalSense_3__habil ;
  vrank: hasRank [ vrank: rankTimestamp "2014-09-01T10:22:57.299892"^^xsd:dateTime ;
    vrank: rankValue "0.143477831349"^^xsd:float ] .

lemon—tmp: translation_5__diestro_right—handed a lemon—tmp: Translation ;
  lemon—tmp: sourceLexicalSense lemon: lexicalSense_2__diestro ;
  lemon—tmp: targetLexicalSense lemon: lexicalSense_6__right—handed ;
  vrank: hasRank [ vrank: rankTimestamp "2014-09-01T10:22:57.300480"^^xsd:dateTime ;
    vrank: rankValue "0.135939716536"^^xsd:float ] .

lemon—tmp: translation_6__able_competente a lemon—tmp: Translation ;
  lemon—tmp: sourceLexicalSense lemon: lexicalSense_8__able ;
  lemon—tmp: targetLexicalSense lemon: lexicalSense_9__competente ;
  vrank: hasRank [ vrank: rankTimestamp "2014-09-01T10:22:57.301556"^^xsd:dateTime ;
    vrank: rankValue "0.0357154492524"^^xsd:float ] .

lemon—tmp: translation_7__able_habil a lemon—tmp: Translation ;
  lemon—tmp: sourceLexicalSense lemon: lexicalSense_8__able ;
  lemon—tmp: targetLexicalSense lemon: lexicalSense_3__habil ;
  vrank: hasRank [ vrank: rankTimestamp "2014-09-01T10:22:57.301883"^^xsd:dateTime ;
    vrank: rankValue "0.0831463214758"^^xsd:float ] .

lemon—tmp: translation_8__able_capaz a lemon—tmp: Translation ;
  lemon—tmp: sourceLexicalSense lemon: lexicalSense_8__able ;
  lemon—tmp: targetLexicalSense lemon: lexicalSense_10__capaz ;
  vrank: hasRank [ vrank: rankTimestamp "2014-09-01T10:22:57.302452"^^xsd:dateTime ;
    vrank: rankValue "0.0357154492524"^^xsd:float ] .

lemon—tmp: translation_9__able_diestro a lemon—tmp: Translation ;
  lemon—tmp: sourceLexicalSense lemon: lexicalSense_8__able ;
  lemon—tmp: targetLexicalSense lemon: lexicalSense_2__diestro ;
  vrank: hasRank [ vrank: rankTimestamp "2014-09-01T10:22:57.302788"^^xsd:dateTime ;
    vrank: rankValue "0.0808980121973"^^xsd:float ] .

```

References

- Blumenstock, J. E. (2008). Size matters: Word count as a measure of quality on Wikipedia. In *Proceedings of the 17th international conference on world wide web, WWW '08* (pp. 1095–1096). ACM, New York, NY, USA. doi:[10.1145/1367497.1367673](https://doi.org/10.1145/1367497.1367673).
- Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, *30*(1–7), 107–117.
- Dwork, C., Kumar, R., Naor, M., & Sivakumar, D. (2001). Rank aggregation methods for the web. *WWW*.
- Fuertes-Olivera, P. A. (2009). The function theory of lexicography and electronic dictionaries: Wiktionary as a prototype of collective free multiple-language internet dictionary. In *Lexicography*

- at a crossroads: *Dictionaries and encyclopedias today, Lexicographical Tools Tomorrow* (pp. 99–134). Bern: Peter Lang.
- Gracia, J., Montiel-Ponsoda, E., Vila-Suero, D., & de Cea, G. A. (2014). Enabling language resources to expose translations as linked data on the web. In *LREC* (pp. 409–413).
- Lih, A. (2004). Wikipedia as participatory journalism: Reliable sources? metrics for evaluating collaborative media as a news resource. In *Proceedings of the 5th international symposium on online journalism* (pp. 16–17). <http://jmsc.hku.hk/faculty/alih/publications/utaustin-2004-wikipedia-rc2.pdf>.
- Lim, E. P., Vuong, B. Q., Lauw, H. W., & Sun, A. (2006). Measuring qualities of articles contributed by online communities. In *Proceedings of the 2006 IEEE/WIC/ACM international conference on web intelligence, WI '06* (pp. 81–87). IEEE Computer Society, Washington, DC, USA. <http://dx.doi.org/10.1109/WI.2006.115>.
- Manola, F., & Miller, E. (2004). *Rdf primer. w3c recommendation*. <http://www.w3.org/TR/2004/REC-rdf-primer-20040210/>.
- Matuschek, M., Meyer, C. M., & Gurevych, I. (2013). Multilingual knowledge in aligned Wiktionary and Omegawiki for translation applications. *Translation: Corpora, Computation, Cognition (TC3)*, 3(1), 87–118.
- McCrae, J., Aguado-de Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gmez-Prez, A., et al. (2012). Interchanging lexical resources on the semantic web. *Language Resources and Evaluation*, 46(4), 701–719.
- 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*, chap. 13 (pp. 259–291). Oxford: Oxford University Press. <http://www.christianmeyer.org/research/publications/oup-elex2012/>.
- Miles, A., & Bechhofer, S. (2009a). *SKOS simple knowledge organization system extension for labels (SKOS-XL)*. <http://www.w3.org/TR/skos-reference/skos-xl.html>.
- Miles, A., & Bechhofer, S. (2009b). *SKOS simple knowledge organization system reference*. <http://www.w3.org/TR/2009/REC-skos-reference-20090818/>.
- Miller, T., & Gurevych, I. (2014). Wordnet—Wikipedia—Wiktionary: Construction of a three-way alignment. In N. C. C. Chair, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.) *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)*. European Language Resources Association (ELRA), Reykjavik, Iceland.
- Montiel-Ponsoda, E., Gracia, J., de Cea, G. A., & Gómez-Pérez, A. (2011). Representing translations on the semantic web. In *MSW* (pp. 25–37).
- Müller, C., & Gurevych, I. (2009). Using Wikipedia and Wiktionary in domain-specific information retrieval. In *Proceedings of the 9th cross-language evaluation forum conference on evaluating systems for multilingual and multimodal information access* (pp. 219–226). CLEF'08 Berlin, Heidelberg: Springer-Verlag.
- Navarro, E., Sajous, F., Gaume, B., Prévot, L., ShuKai, H., & Tzu-Yi, K. et al. (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). People's Web '09 Stroudsburg, PA, USA: Association for Computational Linguistics.
- Sajous, F., Navarro, E., Gaume, B., Prévot, L., & Chudy, Y. (2013). Semi-automatic enrichment of crowdsourced synonymy networks: The wisigoth system applied to Wiktionary. *Language Resources and Evaluation*, 47(1), 63–96.
- Sérasset, G. (2014). DBnary: Wiktionary as a lemon-based multilingual lexical resource in RDF. *Semantic Web Journal: Special issue on Multilingual Linked Open Data*. <http://hal.archives-ouvertes.fr/hal-00953638>.
- Weale, T., & Brew, C. F. L. E. (2009). Using the Wiktionary graph structure for synonym detection. In *Proceedings of the 2009 workshop on the people's web meets NLP: Collaboratively constructed semantic resources, People's Web '09* (pp. 28–31). Association for Computational Linguistics, Stroudsburg, PA, USA. <http://dl.acm.org/citation.cfm?id=1699765.1699769>.
- Zesch, T., Müller, C., & Gurevych, I. (2008). Using Wiktionary for computing semantic relatedness. In *Proceedings of the 23rd national conference on artificial intelligence—Volume 2, AAAI'08* (pp. 861–866). AAAI Press.