Noname manuscript No.
(will be inserted by the editor)

Semi-Automatic Enrichment of Crowdsourced Synonymy Networks: The WISIGOTH system applied to Wiktionary

**Franck Sajous · Emmanuel Navarro ·
Bruno Gaume · Laurent Prévot ·
Yannick Chudy**

Received: date / Accepted: date

Abstract Semantic lexical resources are a mainstay of various Natural Language Processing (NLP) applications. However, comprehensive and reliable resources are rare and not often freely available. Handcrafted resources are too costly for being a general solution while automatically-built resources need to be validated by experts or at least thoroughly evaluated. We propose in this paper a picture of the current situation with regard to lexical resources, their building and their evaluation. We give an in-depth description of Wiktionary, a freely available and collaboratively built multilingual dictionary. Wiktionary is presented here as a promising raw resource for NLP. We propose a semi-automatic approach based on random walks for enriching Wiktionary synonymy network that uses both endogenous and exogenous data. We take advantage of the wiki infrastructure to propose a validation "by crowds". Finally, we present an implementation called WISIGOTH, which supports our approach.

Keywords Synonymy Networks · Semantic Relatedness · Collaboratively Constructed Resources · Wiktionary · Semi-Automatic Enrichment · Random Walks · Small Worlds

1 Introduction

It is a commonplace to underline the importance of lexical resources for Natural Language Processing (NLP) applications. It is also common to complain about their unreliable quality or their prohibitive cost. Many automatic approaches of all sorts have been designed to build such resources but these approaches are not error-free and require human-validation. Such a work is tedious and nevertheless has to be done by experts in order to provide trustworthy resources. However experts' time is precious and relying on them to build handcrafted resources or to validate automatically built ones

---

F. Sajous, B. Gaume and Y. Chudy
CLLE-ERSS, CNRS & Université de Toulouse, France
E-mail: sajous@univ-tlse2.fr

E. Navarro
IRIT, CNRS & Université de Toulouse

L. Prévot
LPL, CNRS & Université de Provence

is too costly. The competitive research context may sometimes be counterproductive: while describing home-made resources and presenting various methods to build them, researchers are reluctant to share these resources. We may also deplore that public investments dedicated to build such resources resulted in poor-quality and non-free ones. Despite various works and an active community, the situation is still not satisfying for most languages. *“We desperately need linguistic resources!”* is claimed by **?**, who argues that it is not realistic to assume that large-scale resources can all be developed by a single institute or a small group of people. Sekine concludes that a collaborative effort is needed, and that sharing resources is crucial.

In this paper, we describe previous attempts to overcome recurrent impediments that hindered the success of resources building. We present new trends based on *crowdsourcing* seen as a promising track to explore (Section 2). We focus then on the problem of automatically extracting synonymy relations. We summarize different existing methods and point out some evaluation problems (Section 3). We study in Section 4 the properties of synonymy networks extracted from reference resources and show that despite sharing a common Hierarchical Small World structure, there are significant discrepancies between them. Such result points out that evaluation in this domain is still an open issue, since gold standards have to be used carefully. We present in Section 5 a free online collaborative dictionary, that could simultaneously settle the problem of cost and (partly) of the evaluation. We present in Section 6 an enrichment process of Wiktionary’s synonymy graphs to reduce their sparseness and measure the impact of using different data sources and similarity measures. We evaluate and comment the results obtained in Section 7. Finally, in Section 8, we present the implementation of our system that everybody can use to improve Wiktionary. We conclude and describe possible extensions of our work in Section 9. We list in Section 10 resources that we extracted from collaborative resources, including Wiktionary, and that can be downloaded.

2 Lexical Resources Building

2.1 Context

Princeton WordNet (**?**), hereafter referred to as WordNet, is probably the only successful and widely used large-scale project among lexical resources building attempts. Following projects EuroWordNet (**?**) and BalkaNet (**?**) were less ambitious in terms of coverage. Moreover, these resources froze when the projects ended while WordNet kept on evolving. EuroWordNet’s problems have been underlined for example in (**?**). Similarly to other methods surveyed in Section 3.3, **?** propose missing relations that require a validation by experts to produce reliable results. Such a validation of the new relations would make the resulting resource very costly and has not been done.
Cost and availability are more and more a matter of concern: in corpus-linguistics, an AGILE-like method borrowed from Computer Science has been proposed by **?** to address the problem of simultaneously maximizing corpus size and annotations while minimizing the time and cost involved in corpus creation. To tackle the availability issue and build free corpora, a method relying on metadata to automatically detect copylefted web pages has been designed by **?**.

In the domain of lexical resources building, methods relying on *crowdsourcing* may help overcoming recurrent bottlenecks.

2.2 Crowdsourcing

Since the birth of Wikipedia, the accuracy of collaboratively constructed resources (CCRs) has been called into question. In the early stages, the only known CCR was the online encyclopaedia and the debate on its accuracy led to a controversy. **?** claimed that the accuracy of the online encyclopaedia comes close to the Britanica encyclopaedia. In return, Britanica criticized the criteria of evaluation (see **?**). From these days, Wikimedia Foundation's projects and other well-known wikis have multiplied. More moderate than Giles, **?** have shown in a task measuring the semantic relatedness of words that resources based on the *"wisdom of crowds"* were not superior to those based on the *"wisdom of linguists"*, but were nevertheless strongly competitive. It has also been demonstrated that crowds can outperform linguists in terms of coverage.

CCRs are clearly better than no resource at all, specially when expert-built resources are out of reach. A problem remains however: how to make people contribute? Indeed, collaborative and social approaches to resource building do not rely only on colleagues or students but on random people that are simply browsing the web and do not share the NLP researchers' interest for linguistic resource building. We enumerate below recent trends for stimulating the crowds to contribure.

Game model Some language resource builders have been successful in designing simple web games to which many people come to play just for fun. For instance, the game *"Jeux de Mots"*[1] developed by **?** has been useful for collecting a great number of relations between words (mostly non-typed associative relations but also better defined lexico-semantic relations such as hypernymy or meronymy). However, setting up a satisfying *gameplay* for collecting any kind of linguistic information is not an easy task. For instance, domain-specific resources might be harder to collect this way. Designing gameplay that really works is a difficult task in itself and it is likely that many initiatives of game-elicited resource will fail because of the game not being fun for the average player.

Mechanical Turk model The Mechanical Turk system has been recently created by Amazon (AMT) and consists in defining *micro-tasks* to be done by workers (*"turkers"*) against a minimal reward (small amount of money or even non monetary reward, such as "reputation"). These tasks are usually impossible or difficult for computers to perform. They are commonly called HITs, for *human intelligence tasks.* Initially, electronic commerce companies used such HITs, for instance, to tag images or to express preference over colors (for a given product). The Wikimedia Foundation used this kind of model to get Wikipedians to rate the articles in order to attribute quality labels. AMT has also been used in the NLP research contexts to overcome the difficulties of carrying out an expert evaluation. For example, **?** used this system to create a collection of Question/Answer sentence pairs. **?** evaluated the performance of non-experts annotation using turkers in natural language tasks such as rating affective text, Word Sense Disambiguation (WSD), word similarity rating, etc. They evaluated

---

[1] See `http://www.lirmm.fr/jeuxdemots/jdm-accueil.php`

these annotations notably by training a supervised system for affect recognition and compared it against the system trained with expert annotations. They obtained the non-intuitive result that for five of seven tasks, the system trained with non-experts annotations outperformed the system trained with the annotations of a single expert. They proposed the explanation that using multiple non-experts may correct the bias of using a single individual labeler. Other experiments led to the conclusion that for many tasks, only a small number of non-experts is necessary to equal the performance of an expert annotator. They found out that an average of four non-expert labels per item provides a score comparable to experts annotation.

tasks?

giving?

AMT is appropriate for some annotation task . However, two constraints put this observations into perspective. The first one relates to human nature: AMT has been designed to perform elementary tasks and should only be used for quick tasks. Otherwise, turkers may be tempted to trick the system by spending a minimal amount of time on each task and give careless answers. Moreover, one person can have many accounts which may reduce the representativeness of the annotator sample. Even in the case of “honest turkers”, task-ability checking may be required depending on the nature of annotations expected (see Section 3.2). The second difficulty is more practical: even if the cost of a task is cheap, it may still be difficult to be funded by a research unit. This can be due to budget shortage or cost being still too expensive or only to administrative complications and unforeseen payment methods (however this calls into question more the functioning of some institutions rather than AMT malfunction).

Piggybacking model Currently, collaborative resources often starts with sophisticated, fancy and costly infrastructures that is waiting for contributors to bring in their knowledge. It is therefore crucial to be popular enough to attract visitor on the platform. Indeed, in the current web landscape, competition for visitors is difficult and empty shells, as promising as they can be, are not attracting many people. Any infrastructure that underestimates and does not answer this attractiveness issue is doomed to fail. Only a few collaborative or social infrastructures are really successful and they concentrate the majority of internet users. Merely being associated with one of these “success stories” affords the possibility of crowds of visitors. Wiktionary and Wikipedia are probably the best examples. The NLP community can offer some services to the users of these resources while taking advantage of their huge amounts of visitors and contributors. Significant steps towards such an architecture have been made in (**??**). Generalizing this approach to social networks, while adding a gaming dimension is also possible and constitutes an interesting avenue to be explored. Moreover, simply adding plugins to existing sound and popular infrastructures requires much less effort and technical skills than setting-up the whole platform (though lots of technical difficulties occur to comply with and plug into these infrastructures).

3 The Case of Synonymy, from the NLP Point of View

Defining linguistic meaning, and in particular modeling synonymy, has been a popular activity among philosophers and theoretical linguists. Giving a synthesis of these works is out of the scope of this paper but we would like to examine the situation in NLP: What kind of synonymy do the applications need? What kind of synonymy are we able to capture? How can we evaluate our models? Indeed, answering a simple question such as *"Are the words* w_1 *and* w_2 *synonymous?"* requires addressing important preliminaries that we introduce below.

3.1 Synonymy Modeling

In (**?**) one can read that *"absolute synonymy, if it exists at all, is quite rare. Absolute synonyms would be able to be substituted one for the other in any context in which their common sense is denoted with no change of truth value, communicative effect or meaning"*. On the same line, **?** states that *"natural language abhor absolute synonyms just as nature abhors a vacuum"*, which is explained by Clark's principle of contrast: even if two words would be absolute synonyms, language works to eliminate them, and either one of the word would fall in disuse or one of them would take a new nuance. So, what kind of synonyms should be included in an NLP semantic resource, and how should them be organized? **?** claims that there is no reason to expect a unique set of word senses can be appropriate for different NLP applications: different corpora can lead to different set of senses and different NLP tasks can require different senses organization. Usually in a resource including synonymy links, two words are synonyms or are not. No further details might be provided. In WordNet, semantic relations organize the synsets, but nothing is said about two lexemes appearing in a same synset. While this situation may be satisfying for some NLP applications, Edmonds and Hirst address the problem of lexical choice in machine translation systems which need to access subtle differences of meaning. To overcome this issue, they propose a model based on a coarse-grained ontology into which clusters of *near-synonyms* represent core meanings. At a fine grain, different kind of contrasts classified into a finite list of variations (denotational, stylistic, expressive, structural, etc.) demarcate the near-synonyms of a given cluster. The discussion of the central role of granularity in this model is very interesting but building a comprehensive lexicon in this way is a huge work and only a small experimental lexicon has been created. Later, **?** proposed methods to automate the building of such a resource. They used the printed *Choose the Right Word* dictionary, which contains clusters of similar words and differences between the words of the same clusters. From this resource, they built a set of clusters (*peripheral concepts* denoting core meanings)[2] that they customized by a mutual bootstrapping process to detect both patterns and pairs of words denoting differences of meanings. Then they added collocation information by processing the *British National Corpus* and using search engine's counts to filter the results. At last, they extracted additional differences of meaning from machine-readable dictionaries. The availability of pre-existing resources is still a strong prerequisite for implementing this method.

Some others authors are relying on mathematical tools to model synonymy: **?** and **?** use maximal cliques to detect word senses in lexical networks. To quote **?**: *"We argue that the various cliques in which a word appears represent different axes of similarity and help to identify the different senses of that word."* However, there is a large discrepancy between lexical networks (see Section 4.2) and the notion of maximal clique is too sensitive to the network chosen: Adding or removing a few links leads to significant differences in the modeling of senses. To address this issue, relying on robust methods, such as the approach proposed by **?**, seems necessary.

---

[2] This process started with OCR scanning, then error correction and annotation.

3.2 The Unresolved Problem of Evaluation

Whatever the model of synonymy chosen for building a resource is, and whatever the target application is, this resource has to be evaluated. Despite numerous attempts, providing a relevant evaluation for synonymy resource is still an open question.

Comparison with gold standards: An usual approach is to evaluate a resource against a gold standard. Provided that such a touchstone exists at all, it is generally not available and if it is, it may not be 100% reliable ; so neither can be the evaluation. Indeed, the resource taken as a gold standard has sometimes been developed for a specific use and cannot cope with an all-purpose evaluation. Therefore, gold standards have to be themselves evaluated or at least characterized before being used for evaluation. It is shown in Section 4.2 that there is not a perfect agreement between gold standards. So, choosing a given gold standard or another may lead to significant differences in evaluation and, therefore, comparing a resource against any gold standard will not permit to draw definitive conclusions. Indeed, whenever a system proposes two words as synonyms which are not synonyms in the gold standard, either the system is wrong or the gold standard is not comprehensive enough.

For example, the method that we developed in (**?**) for enriching the synonymy networks of Wiktionary performed better on the French dictionary than on the English one. Does that mean anything about the initial resources or was it due to the difference of granularity in the French and English gold standards (see Table 11)? Moreover, in (**?**), we explained how we had to adapt our experimental material to comply with gold standards (symmetrizing the edges to evaluate against WordNet and flattening word senses to evaluate against DicoSyn, presented below), which may introduce some bias in the evaluation.

Human evaluation: Evaluating a set of word pairs proposed as synonyms can be done manually by presenting the pairs to human annotators. Unfortunately, this task is subject to high inter-subject variability and often leads to poor inter-tagger agreement (ITA). ITA is frequently presented as the only criterion for quality of a human evaluation. However, even when a satisfying agreement is reached, there is no evidence that the judgments made are good. **?** analyzed the factors correlating with the lack of ITA on a WSD task and found out that high scores are correlated with the annotator's *similarity* (not *level*) of lexical knowledge. Two non-expert judges may obtain the same level of agreement as two experts ; adding an expert to a non-expert team leads to a decrease of ITA. They conclude that agreement alone cannot be taken as a confident measure of correctness but must be combined with some other measure of task ability.

Task-based evaluation: To compare several methods or resources, a common approach is to evaluate the performances of a system using them in a given task. For example, semantic resources may be used in information retrieval (query expansion), machine translation (lexical choice), WSD, detection of near-duplicate contents in documents, etc. To evaluate the system performances, the evaluation process has to determine, for a given input, what output should the system provide. This problem is therefore equivalent to the construction of a gold standard and raises the same problems as stated above. For example, **?** has shown the difficulties of preparing a gold standard for the SENSEVAL competition.

In Section 3.1, we have mentioned the central role of granularity in synonymy modeling. Granularity is crucial in the evaluation process too. In a WSD task evaluation, **?** have shown that grouping the senses of the machine readable dictionary used can reconcile subtle disagreements between annotators. In general, the ITA rose between 10% and 20% when measured against the grouped senses. However, they note that extremely high ITA with highly polysemous words is an unrealistic goal. Moreover, increasing ITA is relevant only if it has no or little impact on NLP applications.

3.3 Discovering Synonymy Relations

In this section we list the main approaches used to collect semantic relations either by relying on corpora, existing lexical networks or even extra-linguistic information.

Pattern-Based Methods: First proposed by **?** to harvest semantic relations from corpora, pattern-based approaches have been refined by **?** by reducing the need of human supervision. Nevertheless human supervision is still necessary and efficient patterns for detecting synonymy are not easy to find when both precision and recall are required. Moreover, such patterns are language-dependent and have to be adapted to other languages. Patterns may however be useful also as a negative filter. Using distributional analysis to detect synonymy relations, **?** applied antonymy patterns to filter potential false positive. If two words (among distributionally similar words) appear often in patterns such as *"from X to Y"* or *"either X or Y"*, they are tagged as antonyms with a 86.4% precision and a 95% recall (and hence removed from detected synonyms).

Vector-Based Models: The most used methods for automatically extracting synonyms consist in building for each word a vector containing various features and to measure similarity between vectors. If two vectors have a high similarity score, the related words are supposed to have a similar meaning. The parameters of these methods are the feature set for the vectors and the similarity measure used. To associate a word with a given vector linguistic features, such as co-occurring words found in corpora may be used, as well as the syntactic contexts. **?** compare bag-of-words and syntactic contexts and study the impacts of linguistic properties (corpus frequency, semantic specificity and semantic classes) on the results. They found out that syntactic contexts outperform bag-of-words and better results are obtained with abstract classes and high-frequency words. The effects of semantic specificity remains unclear. They show also that the extracted relations that are not synonymy are often other semantic relations (co-hyponymy, hypernymy and hyponymy). Comparisons of different measures and weight functions applied on syntactic contexts can be found in (**?**), while **?** examine which particular syntactic context leads to better results. For instance, the *object* relation seems to provide better results than the *adjective* relation.

Cross-Lingual Enrichment of Semantic Resources: Translation links have been used in various wordnets resources to build concepts or to project semantic relations from a language-specific resource to another. **?** used WordNet and bilingual dictionaries to build a Spanish WordNet. They designed and combined different methods to disambiguate bilingual entries against WordNet. Recently, **?** have built WOLF, a *free* French

WordNet, by using several existing resources to bootstrap both concepts (based on synonymy) in French and English and build an inter-lingual index from which resources in each language was able to enrich the other. **?**, for making a proof of concept of language resources interoperability, used translation links to operate an *automatic cross-lingual fertilization* of two lexicons having a WordNet structure. However, all these methods rely on pre-existing lexical resources. Moreover, they produce unavoidable noise and require the human-checking aforementioned which has not been carried on in these experiments.

inversion de aforementioned?

Methods Based on Wiki's Specific Architecture: Different kind of graphs can be built by taking advantage of the specific architectures such as Wikipedia and Wiktionary: for instance, **?** and **?** used the hyperlink structure of the pages or the graph of the article's categories to compute relatedness. Of course, such methods are not reproducible out of these architectures and are not usable with more classical lexical networks.

Random Walks-Based Models: Random walks are efficient methods for computing similarity between vertices of a graph (see for example **??**). Graphs can be built from various data sources: they may model a lexical network into which vertices represent lexemes and edges correspond to semantic relations. Vertices may also be the vectors from the vector-based methods presented above, with edges being weighted by the distance computed between the vectors they link.
We present in Section 6 a method based on random walks over bipartite graphs. We test out both endogenous (synonyms, translations and glosses extracted from Wiktionary) and exogenous (syntactic contexts extracted from a large corpus) data. We also use a bipartite graph mixing these two kinds of data.

4 Properties of Synonymy Networks

In order to account for lexical resources diversity, we have built graphs of synonymy from seven standard French dictionaries (Bailly, Benac, Du Chazaud, Guizot, Lafaye, Larousse and Robert). Synonymy relations have been extracted from each dictionary by the INALF/ATILF Research Unit and corrected by the CRISCO Research Unit. From each of these seven files, we built a non-directed graph. Vertices are lemmas and there is an edge between x and y ($x \longleftrightarrow y$) if and only if x is a synonym of y. We discuss below the structural properties of this kind of graphs (Section 4.1) and then compare them to each others (Section 4.2).

4.1 Invariant Properties of Synonymy Networks

Most of lexical networks, as other Real World Complex Networks (RWCN), are Hierarchical Small Worlds (HSW) networks (**?????**) sharing similar properties. Let $G = (V, E)$ be a symmetric graph, V its set of vertices, and $E \subset V \times V$ its set of edges. We can define:

- $n = |V|$ the order of G (the number of nodes) ;
- $m = |E|$ its size (the number of edges) ;
- $deg(u) = |\{v \in V/(u, v) \in E\}|$ the degree of the node u ;

- $d = \frac{m}{n}$ the average degree.

The four main properties of RWCNs are the following:

- **Edge sparsity:** HSW are sparse in edges ; $m = O(n)$ or $m = O(n\ log(n))$.
- **Short paths:** In HSW, the average path length[3] (L) is short. There is generally at least one short path between any two nodes.
- **High clustering:** In HSW, the clustering coefficient (C) that expresses the probability that two distinct nodes adjacent to a given third one are adjacent, is an order of magnitude higher than for Erdos-Renyi (random) graphs: $C_{HSW} \gg C_{random}$; this indicates that the graph is locally dense, although it is globally sparse.
- **Heavy-tailed degree distribution:** The distribution of the vertices incidence degrees follows a power law in a HSW graph. The probability $P(k)$ that a given node has k neighbours decreases as a power law: $P(k) \approx k^{-\lambda}$ (λ being a constant characteristic of the graph). Conversely, random graphs conform to a Poisson Law.

Table 1 sums-up the structural characteristics of the seven graphs mentioned above. In this table, $\langle k \rangle$ denotes the average degree of the nodes and λ the coefficient of the power law that approximates the distribution of the nodes incidence degrees with a correlation coefficient r^2. When the values are computed on the *largest connected component* they are subscripted by $—_{lcc}$. Other notations are explained above.

Table 1 Structural properties of synonymy graphs.

Dictionnaire	n	m	$\langle k \rangle$	n_{lcc}	C	L_{lcc}	λ	r^2
Bailly	12738	14226	2.38	560	0.04	11.11	−2.67	0.94
Benac	21206	33005	3.33	728	0.02	9.03	−2.68	0.94
Bertaud-du-Chazaud	40818	123576	6.16	259	0.11	6.13	−2.28	0.92
Guizot	3161	2200	2.08	1018	0.08	4.69	−3.56	0.95
Lafaye	3120	2502	2.05	641	0.01	9.37	−2.58	0.97
Larousse	25505	79612	7.11	1533	0.18	6.35	−2.46	0.92
Robert	48898	115763	5.44	3340	0.11	6.43	−2.43	0.94

Even though n and $\langle k \rangle$ vary across dictionaries, L_{lcc} remains low, C is always high, and degrees distribution remains close to a power law ($r^2 > 0.9$) whose coefficient value (λ) is situated between -3.6 and -2.2. This set of properties guarantees that all these networks are HSW.

4.2 Discrepancies Between Synonymy Networks

Although all these graphs are HSW, Table 1 shows that lexical coverage (n) and the number of synonymy links (m) vary significantly across the seven graphs.We therefore focus now on graph comparison.

Given $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, two graphs extracted from the seven dictionaries, we can compute recall, precision and F-score of G_1's lexical coverage against G_2's lexical coverage:

$$R_\bullet(G_1, G_2) = \frac{|V_1 \cap V_2|}{|V_2|}$$

---

[3] Average length of the shortest path between any two nodes.

$$P_{\bullet}(G_1, G_2) = \frac{|V_1 \ \cap \ V_2|}{|V_1|}$$

$$F_{\bullet}(G_1, G_2) = 2.\frac{R_{\bullet}(G_1, G_2).P_{\bullet}(G_1, G_2)}{R_{\bullet}(G_1, G_2) + P_{\bullet}(G_1, G_2)}$$

We notice that $R_{\bullet}(G_1, G_2) = P_{\bullet}(G_2, G_1)$ and that $F_{\bullet}(G_1, G_2) = F_{\bullet}(G_2, G_1)$. $R_{\bullet}(G_1, G_2)$, $P_{\bullet}(G_1, G_2)$ and $F_{\bullet}(G_1, G_2)$ provide information about G_1 and G_2 relative coverage (vertices), but not about their agreement with regard to synonymy (edges). In order to evaluate synonymy links, we must compare the projection of their edges on their shared lexical coverage: $V_1 \cap V_2$. We extract the subgraph $G1_{\langle V_1 \cap V_2 \rangle}$ from G_1 defined as:

$$G1_{\langle V_1 \cap V_2 \rangle} = (V1_{\langle V_1 \cap V_2 \rangle},\ E1_{\langle V_1 \cap V_2 \rangle})$$

where $V1_{\langle V_1 \cap V_2 \rangle} = V_1 \cap V_2$ and $E1_{\langle V_1 \cap V_2 \rangle} = E_1 \cap ((V_1 \cap V_2) \times (V_1 \cap V_2)\ )$.
We define $G2_{\langle V_1 \cap V_2 \rangle}$ in a similar fashion.
To estimate the agreement between G_1 and G_2, we compute recall, precision and F-score of the edges of $G1_{\langle V_1 \cap V_2 \rangle}$ against the edges of $G2_{\langle V_1 \cap V_2 \rangle}$:

$$R_{\updownarrow}(G_1, G_2) = \frac{|E1_{\langle V_1 \cap V_2 \rangle} \ \cap \ E2_{\langle V_1 \cap V_2 \rangle}|}{|E2_{\langle V_1 \cap V_2 \rangle}|}$$

$$P_{\updownarrow}(G_1, G_2) = \frac{|E1_{\langle V_1 \cap V_2 \rangle} \ \cap \ E2_{\langle V_1 \cap V_2 \rangle}|}{|E1_{\langle V_1 \cap V_2 \rangle}|}$$

$$F_{\updownarrow}(G_1, G_2) = 2.\frac{R_{\updownarrow}(G_1, G_2).P_{\updownarrow}(G_1, G_2)}{R_{\updownarrow}(G_1, G_2) + P_{\updownarrow}(G_1, G_2)}$$

Table 2 recaps the evaluation of each pair of graphs as explained above. The agreement on lexical coverage is reported in column ($\bullet$) and the agreement on the synonymy networks restricted to their shared lexical coverage is shown in column ($\updownarrow$). The F-score for edges (boldfaced), ranging from 0.27 to 0.69, with an average value of 0.46, highly depends on the pairs of graphs. This result shows that synonymy, analyzed by expert lexicographers, has a high inter-dictionary variability.

As a consequence of these observations, we merged the seven graphs described above and split this compilation into syntactic categories[4] to obtain three resources: `DicoSyn.Noun`, `DicoSyn.Verb` and `DicoSyn.Adj`. This set of resources will be used as our gold standard for evaluating Wiktionary and our enrichment system in Sections 6 and 7.

5 Wiktionary

We summarize in this section some characteristics of Wiktionary that are relevant for our study. A more comprehensive description of the resource can be found in (**??**).

Wiktionary, the lexical companion to Wikipedia, is a free multilingual dictionary available online. As the other satellites of the Wikimedia Foundation, it is a collaborative project: any user can contribute and its changes are published immediately. Each article may include glosses, etymology, examples, translations and semantic relations

---

[4] The automatic classification into parts of speech and the manual validation has been made at CLLE-ERSS Research Unit by Lydia-Mai Ho-Dac and Franck Sajous.

Table 2 Agreement between pairs of dictionary: Recall (R), Precision (P) and F-Score (F) (to be read row against column.)

		Benac		**Bertaud**		**Guizot**		**Lafaye**		**Larouse**		**Robert**	
		(•)	**(↕)**	**(•)**	**(↕)**	**(•)**	**(↕)**	**(•)**	**(↕)**	**(•)**	**(↕)**	**(•)**	**(↕)**
Bail.	**R**	0.50	0.56	0.29	0.20	0.84	0.60	0.90	0.61	0.40	0.18	0.24	0.20
	P	0.82	0.60	0.93	0.78	0.21	0.49	0.22	0.52	0.81	0.62	0.91	0.71
	F	0.62	**0.58**	0.44	**0.32**	0.34	**0.54**	0.36	**0.56**	0.54	**0.28**	0.37	**0.31**
Ben.	**R**			0.47	0.31	0.85	0.58	0.90	0.68	0.52	0.18	0.30	0.18
	P			0.90	0.76	0.13	0.42	0.13	0.51	0.63	0.60	0.70	0.64
	F			0.62	**0.44**	0.22	**0.49**	0.23	**0.58**	0.57	**0.27**	0.42	**0.28**
Bert.	**R**					0.93	0.78	0.96	0.81	0.76	0.44	0.52	0.54
	P					0.07	0.16	0.07	0.17	0.47	0.38	0.63	0.49
	F					0.13	**0.27**	0.14	**0.29**	0.58	**0.41**	0.57	**0.51**
Guiz.	**R**							0.79	0.68	0.11	0.19	0.06	0.18
	P							0.78	0.69	0.88	0.72	0.91	0.82
	F							0.78	**0.69**	0.19	**0.29**	0.11	**0.29**
Laf.	**R**									0.11	0.18	0.06	0.17
	P									0.93	0.65	0.95	0.77
	F									0.20	**0.28**	0.11	**0.28**
Lar.	**R**											0.44	0.50
	P											0.85	0.54
	F											0.58	**0.52**

such as synonymy/antonymy and hypernymy/hyponymy. For a NLP researcher working on a language for which no reliable resource is freely available, Wiktionary may appear as an Eldorado. The apparent good lexical coverage reinforces this impression. These interesting but impressionistic aspects are completed below by an in-depth study of the resource.

5.1 Information Encoding

Wikicode The Wikimedia Foundation's projects come with a content management system called MediaWiki. A language such as HTML has been judged too difficult to edit for a random contributor and articles' contents are encoded into a language called *wikicode*. Unfortunately, no formal syntax has been defined to specify MediaWiki's wikicode and deviations from the—tacitly agreed?—language are often encountered. We manually analyzed the users' editions and noticed that a non-negligible number of problems in the articles are due to the wiki-syntax misunderstanding or non-compliance. Our intuition is that a significant proportion of users will not become contributors because the wikicode is not so easy to handle.

Entries' layouts A typical article contains potentially several language sections. The first one is the language of the Wiktionay's edition of the article. A language section may start with an etymology section and continue with several part of speech (POS) sections. In a given POS section, we find glosses and examples, sometimes split into different word senses. Then come translations and semantic links.

Unfortunately, there is a lot of variation between the prototypical case and the reality. First, each language has its own conventions. In a given language, the written conventions are not always respected and the last adopted conventions deviate from existing conventions. The notion of *flexibility* is even proclaimed as an intrinsic characteristic of wikis' framework. Consequently, parsing a wiktionary's dump is an uneasy task and things get even worse when dealing with the "history" dump, as both syntax and layouts evolve over time. The practical implications for modeling Wiktionary's synonymy networks are described in (**?**). As a significant illustration, word senses cannot be exploited: The lack of strict format (in Wiktionary) for encoding them prevent their use. Even though glosses can define several word senses in a lexeme section, this sense division is not formally encoded when the senses are used as sources of semantic or translation links. Moreover, targets of semantic or translation relations are lexemes not word senses. Recently, a template has been created to fill this gap but is not used so far (and will probably not be used before long). **?** attempted to disambiguate referred word senses of target lexemes by computing the semantic relatedness between the word sense's gloss of the source and of the senses of the target lexeme. Results were encouraging but limited by the fact that some word senses have too short gloss.

5.2 The Mirage of Numbers

The homepage of the English Wiktionary boasts *"1,758,415 entries with English definitions from over 350 languages"* and the French one *"1,783,911 articles from over 700 languages"*. These impressive numbers have to be tempered. As soon as we look closer, we discover a significant number of meta-articles (help pages, user talks, templates definitions, etc.) that are counted as entries. More surprisingly, a number of foreign entries (regarding the language of the Wiktionary of interest) are included in the count and may represent more than half of the entries announced. Inflected verbal forms or plural forms of nouns are also found as entries when we could expect them inside the related lemmas' articles. Locutions and proverbs occur in Wiktionary but are classified in a strange way and artificially inflate the number of lexemes for a given POS: while *"knowledge is power"* is tagged as a proverb, *"first come first served"* is tagged as a common noun and *"caught between the devil and the deep sea"* is tagged as a standard adjective.[5]

In order to study how evolves a collaborative resource such as Wiktionary, we analyzed the "history dump" provided by the Wikimedia Foundation.[6] This dump contains every version of all articles (stored after each individual contributor's edition) of Wiktionary since its creation (December 2002 for the English edition and March 2004 for the French one). We wrote a parser to index every addition of lexemes[7] and addition/deletion of translation or semantic relations explicitly (formally) encoded. As can be seen in Figure 1, the growth of the English Wiktionary is steady while we notice two jumps in the French edition: the first one (early 2006) is due to an automated import from a public-domain dictionary, the *Dictionnaire de l'Académie Française* (DAF). Other imports have been done gradually, from a second dictionary (*Littré*). Within the English edition, the imports from other dictionaries (mostly *Webster 1913* and *Century 1911*) are not significant. The second jump observed for French (mid-2008), more massive, is due to automated imports of demonyms taken from a specialized web site. This explains why verbs did not undergo this jump. However, one may wonder why

---

[5] These observations are based on March 2010 dumps.

[6] Wiktionaries' dumps are available at: `http://download.wikipedia.org/`

[7] Unfortunately, deleted entries do not occur in the history dump anymore. As a consequence, it is impossible to account for the rate of the lexemes deletion.

Pétrocorien (inhabitant of the town *Périgueux*), together with 76 347 other demonyms, has been included as a standard noun of the dictionary.

Fig. 1 Evolution of the number of lexemes and automated imports in Wiktionary.

In contrast with lexemes, no automated import seems to have been made for synonymy relations. The growth of the semantic relations has been slower than the lexical coverage: Contributors are more prone to add new words than semantic information. When they do, they add mostly synonyms and a few antonyms. Other relations are quite rare. Figure 2 shows the evolution of the semantic links in English and French Wiktionaries. In Figure 3 is depicted the evolution of the number of translation links. No automated import of translation is explicitly mentioned in Wiktionary. Nevertheless, we noticed in the French edition a massive addition of translations (in early 2006) operated by a bot without any explanation. After investigation, we found a very discrete and short discussion in a talk page of the bot's owner stating that his bot automatically added translations taken from an online dictionary without being sure neither if this dictionary has been hand-crafted or checked nor if no copyright prohibits this import.

Fig. 2 Evolution of the number of semantic links in Wiktionary (all POS taken together).

Fig. 3 Evolution of the number of translation links in Wiktionary.

Despite a constant increase of the number of semantic and translation links, the discrepancy between their growth and the growth of the number of lexemes keeps accelerating: see Table 3 for a breakdown of the growth rates between 2007 and 2010.

Table 3 Growth of French and English Wiktionaries from 2007 to 2010.

		2007			**2010**		
		Nouns	**Verbs**	**Adj.**	**Nouns**	**Verbs**	**Adj.**
FR	Lexemes	38 973	6 968	11 787	106 068 (×2.7)	17 782 (×2.6)	41 725 (×3.5)
	Syn.	9 670	1 793	2 522	17 054 (×1.8)	3 158 (×1.8)	4 111 (×1.6)
	Trans.	106 061	43 319	25 066	153 060 (×1.4)	49 859 (×1.2)	32 949 (×1.3)
EN	Lexemes	65 078	10 453	17 340	196 790 (×3.0)	67 649 (×6.5)	48 930 (×2.8)
	Syn.	12 271	3 621	4 483	28 193 (×2.3)	8 602 (×2.4)	9 574 (×2.1)
	Trans.	172 158	37 405	34 338	277 453 (×1.6)	70 271 (×1.9)	54 789 (×1.6)

5.3 Size of Headword List and Lexical Coverage

Despite the automated imports of demonyms and some other questionable choices, the size of Wiktionary's headword list looks more than respectable. We wanted to check how much Wiktionary overlaps with more traditional dictionaries. We compared the lexemes contained in the French collaborative resource with the *Trésor de la Langue Française informatisé* (TLFi), an handcrafted dictionary developed at the INALF/ATILF Research Unit by expert lexicographers. The TLFi's headword list has been extracted from a freely available lexicon called Morphalou[8]. Table 4 shows that Wiktionary contains 3/4 of the TLFi's nouns, almost all its verbs and 2/3 of its adjectives. In order to evaluate to what extent Wiktionary could be used as a resource for NLP, we extracted the vocabulary from 3 different corpora: *Frantext20* is a 30 million words corpus including 515 novels from the 20th century ; *LM10* is a 200 million words corpus containing the articles of the daily newspaper *Le Monde* over a 10 year period ; *Wikipedia2008* is a 260 million words corpus extracted from the French Wikipedia in year 2008. Each corpus has been tagged and lemmatized with TreeTagger. Then we built for each corpus a list of lemmas with a frequency greater than 4 and we observed how much the headword list of the TLFi and Wiktionary overlap with the corpora's vocabularies. For both dictionaries, the coverage is better on Frantext20 than LM10 and better on LM10 than Wikipedia. The low coverage on Wikipedia may be due to the wide range of contributors and topics, as well as tokenization problems and a significant number of words from different languages. The lowest coverage for Wikipedia's nouns may be explained by a large number of isolated words unknown to TreeTagger often tagged as nouns. Wiktionary has always a better coverage for nouns and verbs (2% to 7%) and the TLF has a better coverage for the adjectives (1% to 4%). Building the intersection of the headword lists (refered to as T∪W) leads to a rise of coverage for nouns (5%) and adjectives (10%). These results show that despite the noisy nature of Wiktionary, it is worth building NLP resources from it for text analysis tasks. These results also confirm the observations made in (**?**): crowdsourced resources and expert-built resources do not overlap exactly but contain complementary knowledge. Indeed, Wiktionary does not only contain neologisms taken from the Internet field such as *googler* (to google) and *wikifier* (to wikify). It contains also domain-specific words such as *cryosphère* (cryosphere) or *clitique* (clitic) and words that have now become part of standard usage such as *societal* (societal), *ergonomique* (ergonomic), *décélérer* (to decelerate), *étanchéifier* (to waterproof), *paramétrer* (to parameterize), etc.

word ?

years ?

word ?

with/having ?

Table 4 Wiktionary (2011) and the TLFi's Lexical Coverages.

	Size of the headword list			**% of Lexival Coverage Regarding Corpora**								
				Frantext20			**LM10**			**Wikipedia2008**		
	TLFi	Wikt.	Intersection	TLFi	Wikt.	T∪W	TLFi	Wikt.	T∪W	TLFi	Wikt.	T∪W
N.	41005	134203	29604	76,4	80,6	84.4	47,3	54,1	58,1	23,5	26,7	31,6
V.	7384	18830	6964	84,2	86,5	87.1	75,1	80,0	80,8	66,3	71,5	72,2
Adj.	15208	42263	10014	88,9	84,6	94.0	78,9	76,8	88,1	73,9	72,4	84,7

6 Semi-Automatic Enrichment of Wiktionary

Based on the fact that resources extracted from Wiktionary are very sparse with regard to synonymy relations (cf. Table 3), we made an attempt in (**?**) to enrich it. Relying only on endogenous data (i.e. the existing synonymy links), we used *Prox*, a

---

[8] `http://www.cnrtl.fr/lexiques/morphalou/`

stochastic method presented in (**?**) for computing a similarity measure between two nodes (lexemes). We proposed to connect each vertex u to the k first vertices ranked in descending order with respect to the *Prox* measure, k being chosen proportionally to the original incidence degree (number of neighbors) of u. We compared the resource obtained after this enrichment to gold standards. We observed unsurprisingly that adding a small amount of links leads to a poor gain of recall and a small decrease of precision, while adding a large amount of links significantly increases the recall and decreases the precision. However we significantly improved the F-score. For instance, we managed to double the number of synonymy links for French verbs with only a 2% loss of precision. This evaluation method suffered from the bias of using a gold standard, as discussed in Section 3.2. Moreover, producing a reliable resource would require a human-validation, which, as stated in Section 2, may be unaffordable.

We decided for the current work, as described hereafter, to use a comparable approach to compute the candidate synonymy relations. An innovation compared to the previous method consists in adapting this approach in a perspective of collaborative editing: We introduce now a validation process intended to be performed by Wiktionary's contributors. Hence, an automatically computed candidate synonymy relation is suggested to contributors that can decide whether this relation has to be added to Wiktionary or not.
This approach sorts out the problem of validation (apart from the question of the lexical knowledge of these contributors). Another question remain however: Choosing the number of neighbors to be added to a given lexeme proportionally to its original incidence degree seems “fair” but might be problematic. Indeed, in a collaborative resource, if a lexeme has few synonyms, one cannot decide whether it does reflect the reality (low polysemy) or it stems from contributors not having yet worked on the corresponding entry. Relying on a “crowds-based” validation assumes contributors will choose a relevant number of neighbors depending on their nature and the candidates being proposed.

Another innovation consists in adding exogenous data to endogenous ones considered so far. We study below the impact of using several data sources and different similarity measures.

6.1 Weighted Bipartite Graphs Model

In order to homogenize and simplify the description of experiments, each type of data we used is modeled as a *weighted undirected bipartite graph* $G = (V, V', E, w)$ where the set of vertices (V) always corresponds to the lexemes of the language and part of speech of interest, whereas another set of vertices (V') varies according to the data source. The set of edges (E) is such that $E \subseteq (V \times V')$. It models the relations between the lexemes of V and V'. Moreover, a weight is given to each edge by the function $w : E \to \mathbb{R}^+$.

Translations graph $G_{Wt} = (V, V_{Wt}, E_{Wt}, w_{Wt})$
Here, $V' = V_{Wt}$ is the set of the lexemes in all languages but the one of interest. E_{Wt} is the set of translation links: There is an edge between $v \in V$ and $t \in V_{Wt}$

if t is found as a translation of v.[9] There is no particular weight on the edges, so $\forall e \in E,\ w_{Wt}(e) = 1$.

Synonyms graph $G_{Ws} = (V, V_{Ws}, E_{Ws}, w_{Ws})$

Here, $V' = V_{Ws}$ is simply a copy of V. There is an edge between $v \in V$ and $u \in V_{Ws}$ when $v = u$ or u (or v) is indicated as synonym in v entry (or u entry). Similarly to translation graph, there is no particular weight on the edges: $\forall e \in E,\ w_{Ws}(e) = 1$. This bipartite graph model of the synonymy network may look unusual, however: (i) it permits to have a unique bipartite graph model, (ii) for the random walk algorithms presented below, this model is equivalent to a classic unipartite synonymy network.

Glosses graph $G_{Wg} = (V, V_{Wg}, E_{Wg}, w_{Wg})$

Here, $V' = V_{Wg}$ corresponds to the set of all lemmatized lexemes found in the glosses of all entries. There is an edge between $v \in V$ and $g \in V_{Wg}$ if g is used in one of the definitions of v. For a given lexeme, glosses have been concatenated, lemmatized, tagged with Treetagger,[10] and stopwords have been removed. Various weights may be used here but we simply used frequency. The weight of the edge between $u \in V$ and $g \in V_{Wg}$ is the number of occurrences of g in u's gloss. Note that the position in the gloss may also be a relevant weighting factor.

Graph of Wikipedia's syntactic contexts $G_{Wpc} = (V, V_{Wpc}, E_{Wpc}, w_{Wpc})$

We extracted a 260 million words corpus from the French Wikipedia and analyzed it with Syntex, a syntactic parser for French (**?**). This parser produces dependency relations that we used to construct a list of syntactic cooccurrents by building up a frequency table of `<lexeme,context>` pairs, the context consisting of another lexeme and a syntactic relation linking both lexemes (e.g. how many times noun N occurs as an *object* of verb V). V_{Wpc} is the set of syntactic contexts and there is an edge $e = (v, c) \in E_{Wpc}$ as soon as the lexeme v appears in context c. We used pointwise mutual information to weight these edges:

$$\forall (v,c) \in E, w_{Wpc}((v,c)) = \log(\frac{f(v,c)f(*,*)}{f(v,*)f(*,c)})$$

where $f(v,c)$ is the frequency of the lexeme v in the context c, $f(v,*)$, $f(*,x)$ and $f(*,*)$ are respectively the total frequency of v (within any context), the total frequency of c (with any lexeme) and the total frequency of any pair.

Graphs merging We used different combinations of the graphs introduced above, as can be seen in Table 6 presented with their respective order and size. For example "$s + t + g$" is the graph containing synonymy, translation and glosses links, or, more formally:

$$G = \big(V, \quad V' = V_{Ws} \cup V_{Wt} \cup V_{Wg}, \quad E = E_{Wt} \cup E_{Wt} \cup E_{Wg}, \quad w\big)$$

Note that two vertices from different "V'" (for example one in V_{Wt} and one in V_{Wg}) are always considered as dissimilar even if they correspond to the same lexeme. We weight these graphs by multiplying edges' weights by a positive coefficient in function

---

[9] As we parse only the dump of the language of interest, we find the *oriented* link $v \rightarrow t$ (t as a translation of the lexeme v in v's entry) and made it symmetric into $v \leftrightarrow t$. Having a more subtle model (including oriented edges) would require the ability to parse all dumps of all languages.

[10] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

Table 5 Order and size of the bipartite graphs used to compute candidate synonyms. n and n' are the number of vertices, respectively in V and V', which count at least one neighbor. m is the number of edges.

		English			**French**		
		n	n'	m	n	n'	m
A.	trans	8178	43976	54840	5335	23976	32944
	syn	8723	8723	27257	4482	4482	12754
	glosses	45703	39409	218993	41620	42455	263281
	contexts	–	–	–	6262	129199	934969
V.	trans	7473	52862	70432	3174	30162	49866
	syn	7341	7341	23927	3190	3190	9510
	glosses	42901	36051	222004	17743	16942	101458
	contexts	–	–	–	4273	2312096	5499611
N.	trans	29489	235233	277897	18468	129426	153033
	syn	31227	31227	86195	19407	19407	53869
	glosses	194694	127198	1218414	105760	69994	844805
	contexts	–	–	–	22711	1671655	8719464

of the edges' types. The graph denoted "$\alpha_s.s\ +\ \alpha_t.t\ +\ \alpha_g.g$" will have the following weighting function:

$$w(e) = \begin{cases} \alpha_s.w_{Ws}(e) & \text{if} \quad e \in E_{Ws}, \\ \alpha_t.w_{Wt}(e) & \text{if} \quad e \in E_{Wt}, \\ \alpha_g.w_{Wg}(e) & \text{if} \quad e \in E_{Wg}. \end{cases}$$

This is clearly not the only way neither to weight such a combined graph nor to aggregate such data sources. For instance, we could have first computed the lists of candidates for each data source and then aggregated it. It is nevertheless a simple method which permitted to significantly increase the number of relevant candidates proposed by the system (see evaluations in Section 7.2).

Table 6 Order and size of the bipartite graphs combinations used to compute candidate synonyms. n and n' are the number of vertices, respectively in V and V', which count at least one neighbor. m is the number of edges. "s" means synonyms graph, "t" translations graph, "g" glosses graph and "c" Wikipedia's syntactic contexts graph.

		English			**French**		
		n	n'	m	n	n'	m
A.	s + t	13650	52699	82097	7849	28458	45698
	s + t + g	47280	92108	301090	42507	70913	308979
	s + t + g + c	-	-	-	42517	200761	1248779
V.	s + t	11423	60203	94359	5054	33352	59376
	s + t + g	44295	96254	316363	18226	50294	160834
	s + t + g + c	-	-	-	18229	2374679	5700602
N.	s + t	50305	266460	364092	30810	148833	206902
	s + t + g	202920	393658	1582506	111228	218827	1051707
	s + t + g + c	-	-	-	111290	1898564	9818553

6.2 Random Walk-Based Similarity Computation

To propose new synonymy relations, we compute the similarity between any possible pair of lexemes (the vertices from the graphs described in the previous section). The objective is to propose the pairs with the highest scores as candidates for synonyms (which are not already known as synonyms in Wiktionary). We test various similarity measures, all based on short fixed length random walks. Such approaches are introduced in (**??**) for measuring topological resemblance in graphs. This kind of methods has also been applied to lexical networks in (**?**) to compute semantic relatedness. We consider a walker wandering at random along the edges of the *weighted undirected bipartite graph* $G = (V \cup V', E, w)$ and starting from a given vertex v. At each step, the probability for the walker to move from nodes i to j is given by the cell (i, j) of the transition matrix P, defined as follow:

$$[P]_{ij} = \begin{cases} \frac{w((i,j))}{\sum_{k \in \mathcal{N}(i)} w((i,k))} & \text{if} \quad (i,j) \in E, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

where $\mathcal{N}(i)$ is the set of neighbors of the vertex i: $\mathcal{N}(i) = \{j/(i,j) \in E\}$. Thus, starting from v, the walker's position after t steps is given by the distribution of probabilities $X_t(v) = \delta_v P^t$, where δ_v is a row vector of dimension $|V \cup V'|$ with 0 anywhere except 1 for the column corresponding to vertex v. We note $X_t(v, u)$ the value of the coordinate u of this vector, which denotes as aforementioned the probability of reaching u after t steps, starting from v. This is the first measure[11] (called *simple*) we use ; other measures are based on this one:

$$\text{simple}(v,u) = X_t(v,u) \tag{2}$$

$$\text{avg}(v,u) = \frac{X_t(v,u) + X_t(u,v)}{2} \tag{3}$$

$$\cos(v,u) = \frac{\sum_{w \in V} X_t(v,w) X_t(u,w)}{\sqrt{\sum_{w \in V} X_t(v,w)^2} \sqrt{\sum_{w \in V} X_t(u,w)^2}} \tag{4}$$

$$\text{dot}(v,u) = \sum_{w \in V} X_t(v,w) X_t(u,w) \tag{5}$$

$$\text{ZKL}_\gamma(v,u) = \sum_{w \in V} X_t(v,w) \begin{cases} \log(\frac{X_t(v,w)}{X_t(u,w)}) \text{ if } X_t(u,w) \neq 0 \\ \gamma \text{ otherwise} \end{cases} \tag{6}$$

"cos" and *"dot"* are respectively the classical cosine and scalar product. "ZKL_γ" is a variant of the Kullback-Leibler divergence introduced by **?**.

Let $C(v, G, t, \text{sim})$ be the ordered list of candidates computed on graph G with the similarity measure *"sim"* and a random walk of length t, starting from v:

$$C(v, G, t, \text{sim}) = [u_1, u_2, \dots, u_n] \quad \text{with} \quad \begin{cases} \forall i, \text{ sim}(v, u_i) \geq \text{sim}(v, u_{i+1}) \\ \forall i, \text{ sim}(v, u_i) > 0 \\ \forall i, \ (v, u_i) \notin E_{Ws} \end{cases} \tag{7}$$

where E_{Ws} is the set of existing synonymy links in Wiktionary.

The experiments below consist in evaluating the relevancy of $C(v, G, t, \text{sim})$ when G, and sim vary, whereas $t = 2$ remains constant.[12]

---

[11] All these measures are not strictly speaking *similarity*, indeed "simple" and "zkl10" are not symmetric.

[12] Indeed, t has to be even and preliminary experiments have shown that best results are obtained with $t = 2$ or $t = 4$, when $t = 2$ gives similar results and is less complex.

7 Evaluation

7.1 Evaluation Method

With our application in mind (cf. Section 8.2) and given the principle of a semi-automatic approach in which contributors select the candidates to be added, we consider for each lexeme that a suggested list of candidates is *acceptable* when it includes at least one relevant candidate. Indeed, an user can contribute provided that at least one good candidate occurs in the suggested list. Thus, the evaluation will broadly consist in counting for how many lexemes the system computes a suggested list with at least one relevant candidate. Nevertheless we also count how many lexemes have 2, 3 or more good candidates.

Let $G_{GS} = (V_{GS}, E_{GS})$ be a gold standard synonymy network, where V_{GS} is a set of lexemes, and $E_{GS} \subseteq V_{GS} \times V_{GS}$ a set of synonymy links. We evaluate below the acceptability of the suggested lists made to enhance the deficient resource against the gold standard's relations. We only evaluate the suggested lists for the lexemes that are included in the gold standard (i.e. $v \in V_{GS}$). In cases where a lexeme $v \in V$ does not belong to the gold standard (i.e. $v \notin V \cap V_{GS}$), we consider it as a lexical coverage issue. As a result we cannot deem whether a relation (v, c) is correct or not.[13] For the same reason, for each lexeme v, we remove from $C(v)$ the candidates that were absent from the gold standard. Finally we limit the maximum number of candidates to $k \leq 5$. For each lexeme $v \in V \cap V_{GS}$, we note $\mathit{\Gamma}_k(v)$ the "evaluable" suggested list of candidates:

$$\mathit{\Gamma}_k(v) = [c_1, c_2, \ldots, c_{k'}] \quad \text{with} \quad \begin{cases} k' \leq k \\ \forall i,\ c_i \in C(v) \cap V_{GS} \\ \forall i, \mathrm{sim}(v, c_i) \geq \mathrm{sim}(v, c_{i+1}) \end{cases} \tag{8}$$

Please note that $\mathit{\Gamma}_k(v)$ contains a maximum of k candidates (but it may be smaller or even empty). Note also that $\mathit{\Gamma}_k(v)$ depends on the gold standard. We note $\mathit{\Gamma}_k^+(v)$ the set of correct candidates within $\mathit{\Gamma}_k(v)$:

$$\mathit{\Gamma}_k^+(v) = \left\{ c^+ \in \mathit{\Gamma}_k(v) \big/ (v, c^+) \in E_{GS} \right\} \tag{9}$$

We define the set N_k of lexemes having k candidates being proposed and the subset N_k^{+p} of lexemes for which at least p *correct candidates* are proposed:[14]

$$N_k = \left\{ v \in V \cap V_{GS} \big/ |\mathit{\Gamma}_k(v)| = k \right\} \ , \ N_k^{+p} = \left\{ v \in N_k \big/ |\mathit{\Gamma}_k^+(v)| \geq p \right\} \tag{10}$$

To compare the virtues of different data sources for computing the candidates, we measure R_k, the ratio between the number of suggested lists and the number of evaluable target lexemes, and P_k, the ratio between the *acceptable* suggested lists (i.e. lists counting at least one good candidate) and the lexemes for which suggestions are made:

---

[13] v may be a neologism or a domain-specific word. Less often, it may be a misspelling. Any relation (v, c) should therefore not be counted as good (or wrong).

[14] Definitions of N_k and N_k^{+p} differ from those used in (?). These sets are here limited to lexemes for which are proposed at least k evaluable candidates instead of at least one in the previous proposal. The reason is that lexemes for which only one candidate is proposed have lower chances to find a correct candidate and there is no chance to find two correct ones. So considering lexemes that have only one candidate creates a negative bias in the measure.

$$R_k = \frac{|N_k|}{|V_{GS} \cap V|}, \qquad P_k = \frac{|N_k^{+1}|}{|N_k|} \tag{11}$$

Although P_k and R_k are not precision and recall measures, they intuitively refer to the same notions and we adopt below—abusively—this terminology.

Gold Standards: We used WordNet to evaluate the candidates for English and DicoSyn (see Section 4.2) for French. The extraction of the synonymy networks from these resources reproduces what has been done in (**?**). The size and properties of these graphs are presented in Table 7.

Table 7 Properties of the gold standard's synonymy graphs.

Graph	n	m	n_{lcc}	m_{lcc}	$\langle k \rangle_{lcc}$	L_{lcc}	C_{lcc}	λ_{lcc}	r^2_{lcc}
PWN.Noun	117798	168704	40359	95439	4.73	7.79	0.72	-2.41	0.91
PWN.Adj	21479	46614	15945	43925	5.51	6.23	0.78	-2.09	0.9
PWN.Verb	11529	40919	9674	39459	8.16	4.66	0.64	-2.06	0.91
DicoSyn.Noun	29372	100759	26143	98627	7.55	5.37	0.35	-2.17	0.92
DicoSyn.Adj	9452	42403	8451	41753	9.88	4.7	0.37	-1.92	0.92
DicoSyn.Verb	9147	51423	8993	51333	11.42	4.2	0.41	-1.88	0.91

7.2 Results

Similarity measures: Applying the similarity measures presented in Section 6.2 leads to pretty comparable results. For instance, the results obtained with the synonyms graph, translations graph and the union of this two graphs for the English and French Wiktionaries' nouns and verbs are reported in Table 8. Since the *simple* measure is as efficient as the others while being much simpler (quicker computable), further experiments have been done using this measure.

Table 8 P_5 Precision comparison for different data sources and measures.

	Synonyms				**Translations**				**Syn. + Trans.**			
	EN		FR		EN		FR		EN		FR	
	V	N	V	N	V	N	V	N	V	N	V	N
simple	44.5	34.6	68.0	54.4	**60.4**	51.0	**90.4**	**79.7**	**58.0**	47.6	**85.6**	**66.9**
avg	46.1	36.8	**68.7**	54.5	58.9	51.2	**90.4**	78.9	57.0	48.0	84.7	66.5
cos	**46.5**	**37.7**	66.7	55.1	58.8	50.9	90.1	78.5	56.4	48.0	84.4	65.3
dot	45.3	36.7	66.0	53.1	59.7	**51.4**	90.1	79.0	57.7	**48.5**	84.7	66.4
ZKL$_{10}$	46.4	37.1	66.3	**55.4**	58.3	50.8	88.8	78.1	56.8	48.3	83.9	65.4

Data sources: As we can see in Tables 9 and 10, better results are obtained for French than for English. As discussed in Section 3.2, this can be partly explained by the slightly lower density of the English networks (cf. Table 3). However it is mainly due to the difference between the gold standards used: Networks extracted from WordNet are sparser than the ones extracted from DicoSyn (cf. Table 7) that is the result of merging seven graphs extracted from seven dictionaries (see Section 4.2).

Table 9 Impact of different data sources on the *simple* similarity measure. N_5 is the set of lexemes having k candidates being proposed, N_5^{+p} is the set of lexemes for which at least p correct candidates are proposed. *nw*-gloss is an unweighted version of the glosses graph.

			R_5	P_5	$\|N_5\|$	$\|N_5^{+1}\|$	$\|N_5^{+2}\|$	$\|N_5^{+3}\|$	$\|N_5^{+4}\|$	$\|N_5^{+5}\|$
EN	**Adj.**	syn	17.4	49.1	2456	1207	439	**165**	**57**	**22**
		trans	9.2	**65.7**	1299	853	406	144	27	3
		nw-glosses	**93.5**	25.9	**13205**	3421	774	154	34	2
		glosses	**93.5**	26.6	**13205**	**3510**	**794**	158	30	1
	Nouns	syn	8.7	34.6	3862	1335	483	**200**	**95**	**54**
		trans	8.5	**51.0**	3759	1916	655	178	41	2
		nw-glosses	**95.6**	14.8	**42337**	6252	926	106	6	0
		glosses	**95.6**	15.3	**42337**	**6467**	**933**	114	5	1
	Verbs	syn	23.9	44.5	2153	959	431	**216**	**115**	**59**
		trans	24.7	**60.4**	2223	1342	**609**	187	43	1
		nw-glosses	**98.5**	27.0	**8852**	2389	518	98	10	2
		glosses	**98.5**	28.1	**8852**	**2490**	548	100	13	2
FR	**Adj.**	syn	11.9	75.2	480	361	224	**139**	55	**16**
		trans	6.0	**91.4**	243	222	184	117	**56**	11
		nw-glosses	**90.2**	32.2	**3627**	1167	309	91	12	1
		glosses	**90.2**	33.6	**3627**	**1220**	**337**	100	17	0
		contexts	86.2	20.7	3468	719	157	40	11	1
	Nouns	syn	10.4	54.4	1722	936	478	194	68	15
		trans	5.5	**79.7**	916	730	472	**245**	**94**	**20**
		nw-glosses	**95.8**	20.6	**15828**	3268	607	116	16	2
		glosses	**95.8**	22.5	**15828**	**3560**	693	127	21	3
		contexts	84.0	20.9	13882	2898	**721**	181	34	5
	Verbs	syn	10.0	68.0	412	280	172	86	30	5
		trans	19.0	**90.4**	785	710	544	**352**	**146**	**38**
		nw-glosses	**95.6**	41.2	**3947**	1628	530	149	38	3
		glosses	**95.6**	44.9	**3947**	**1773**	**638**	198	45	8
		contexts	81.8	35.3	3378	1192	426	126	28	3

The translations graph provides better precision than synonymy graphs. This result was expected since in Wiktionary, lexemes have more translation links than synonyms (see Table 5). Moreover, translations are often distributed over several languages, which is more reliable than having a lot of translations into a unique language. Using Wiktionary's glosses and Wikipedia's contexts provided unsurprisingly the worse precision and highest recall: Almost all lexemes have glosses in the dictionary and occur in the corpus, but information is less specific. Note that using the lexemes' frequency to weight the graphs of glosses slightly improves the results. A more tricky weighting (for example, by favoring the lexemes occurring at initial positions in the glosses) may perform even better. Curiously, Wikipedia's syntactic contexts lead to a quite poor result in terms of precision, which is opposite to the results found in the literature (e.g. **?**). Certainly filtering rare contexts (with a simple frequency threshold) should improve this result. When lexemes occur only with a single syntactic context, they tend to have a high mutual information without being really significant for bringing closer the lexeme to another one occurring with the same context.

Results using combined data sources are given in Table 10. Combining synonyms and translations enables a better recall than with separated graphs and a similar precision for English. In French resources, it leads to a loss of precision compared to the "translations only" graph. As soon as glosses are used, candidates may be proposed for almost all lexemes ($R_5 \geq 90\%$). The better precision is obtained by weighting synonyms and translations ten times more than glosses, and for French, glosses again ten times more than syntactic contexts (i.e. graphs “$10.s + 10.t + g$” for English and “$10^3.s + 10^3.t + 10^2.g + c$” for French). Using these last graphs enables us to propose a list of 5 candidates for almost all lexemes and between 35% and 60% of these lists count at least one candidate validated by a gold standard.

Table 10 Impact of combined data sources on the *simple* similarity measure. N_5 is the set of lexemes having k candidates being proposed, N_5^{+p} is the set of lexemes for which at least p correct candidates are proposed. Graphs names have the following meaning: “s”: Synonyms graph, “t”: Translations graph, “g”: Glosses graph, “c”: Wikipedia’s syntactic contexts graph.

			R_5	P_5	$\|N_5\|$	$\|N_5^{+1}\|$	$\|N_5^{+2}\|$	$\|N_5^{+3}\|$	$\|N_5^{+4}\|$	$\|N_5^{+5}\|$
EN	**A.**	$s + t$	21.9	**58.1**	3096	1800	805	283	91	29
		$\mathbf{10.s + 10.t + g}$	**95.0**	35.9	**13417**	**4819**	**1567**	**455**	**125**	**32**
		$\mathbf{10^2.s + 10^2.t + g}$	**95.0**	35.9	**13417**	**4818**	**1567**	**455**	**125**	**32**
	N.	$s + t$	14.5	**47.6**	6440	3063	1061	348	110	45
		$\mathbf{10.s + 10.t + g}$	**96.4**	23.3	**42688**	**9944**	2344	**561**	**142**	**43**
		$\mathbf{10^2.s + 10^2.t + g}$	**96.4**	23.3	**42688**	9942	**2345**	**561**	**143**	**43**
	V.	$s + t$	37.6	**58.0**	3380	1962	918	358	119	**43**
		$\mathbf{10.s + 10.t + g}$	**99.2**	41.0	8916	**3655**	**1352**	**448**	**136**	34
		$10^2.s + 10^2.t + g$	**99.2**	40.9	**8917**	3644	1351	**448**	**136**	34
FR	**A.**	$s + t$	15.7	**81.3**	631	513	375	243	105	28
		$10.s + 10.t + g$	89.5	44.3	3602	1594	728	371	154	38
		$10^2.s + 10^2.t + g$	91.2	43.6	3668	1600	729	370	155	38
		$10^2.s + 10^2.t + 10.g + c$	**97.3**	41.9	3913	1640	680	347	143	32
		$\mathbf{10^3.s + 10^3.t + 10^2.g + c}$	**97.3**	45.3	**3915**	**1774**	**791**	**408**	**172**	**43**
	N.	$s + t$	15.2	**66.9**	2511	1681	983	480	166	33
		$10.s + 10.t + g$	96.5	33.3	15948	5303	1956	735	219	50
		$10^2.s + 10^2.t + g$	96.5	33.2	15948	5298	1952	736	218	52
		$10^2.s + 10^2.t + 10.g + c$	**98.5**	33.1	**16274**	5394	1908	649	196	38
		$\mathbf{10^3.s + 10^3.t + 10^2.g + c}$	98.4	36.7	16273	**5980**	**2240**	**825**	**260**	**56**
	V.	$s + t$	25.7	**85.6**	1062	909	669	418	165	48
		$10.s + 10.t + g$	96.6	55.9	3989	2229	1161	580	216	**58**
		$10^2.s + 10^2.t + g$	96.6	55.8	3989	2226	1160	580	214	**58**
		$10^2.s + 10^2.t + 10.g + c$	**98.1**	53.2	**4053**	2158	1004	433	146	43
		$\mathbf{10^3.s + 10^3.t + 10^2.g + c}$	**98.1**	58.4	**4053**	**2368**	**1243**	**604**	**223**	53

Table 11 Example of candidate synonymy relations for nouns evaluated against gold standards (GS). All lexemes belong to both Wiktionary and the gold standards.

	in GS	**Propositions**
EN	Yes	*<imprisonment: captivity>*, *<harmony: peace>*, *<filth: dirt>*, *<antipasto: starter>*, *<load: burden>*, *<possessive: genitive>*, *<stem: radical>*, *<fellow: colleague>*, *<underworld: Hell>*, *<neighborhood: neighbourhood>*, *<words: quarrel>*, *<words: speech>*
	No	*<rebirth: renewal>*, *<fool: idiot, dummy>*, *<cheating: fraud>*, *<bypass: circumvention>*, *<dissimilarity: variance>*, *<pro: benefit>*, *<dog: bitch>*, *<hound: greyhound>*, *<taste: flavour>*, *<inaccuracy: inexactitude>*, *<store: warehouse>*, *<belongings: possession>*
FR	Yes	*<ouvrage, travail>*, *<renom: gloire>*, *<emploi: fonction>*, *<drapeau: pavillon>*, *<rythme: cadence>*, *<roulotte: caravane>*, *<chinois: tamis>*, *<contribution: cotisation>*, *<bobard: tromperie>*, *<cabinet: chiotte>*, *<soupe: bouillon>*, *<nombre: effectif>*
	No	*<drogue: psychotrope>*, *<fantassin: bidasse>*, *<force: poigne>*, *<salade: bobard>*, *<W.C.: chiotte>*, *<us: tradition>*, *<dico: lexique>*, *<job: emploi>*, *<taf: profession>*, *<cantoche: cantine>*, *<souscription: cotisation>*, *<bisque: soupe>*, *<nombre: valeur>*

Table 11 shows some examples of candidates computed by the enrichment process using the $s + t$ graph (combination of synonyms and translations). Some of them are close synonyms (*possessive, genitive*), some others are just geographical variants – different lexical unit (*gas station, petrol station*) or different spelling (*neighbourhood, neighborhood*). Several candidates for the same target word may denote several senses of this word (*words/quarrel* and *words/speech*). By evaluating these candidates against gold standards, we can notice that some rejected propositions seem quite reasonable.

Some computed pairs are linked in the gold standard by hypernymy/hyponymy relations (*hound, greyhound*). Some oppositions between positive and negative judgments show the limits of the evaluation against gold standards, which makes it hard to draw definitive conclusions. Indeed, *cabinet* is a synonym of *chiotte*[15], but it is unclear why *W.C.* is not. It is also interesting to notice the impact of using gold standards of different kinds: in WordNet, which contains both synonymy and hypernymy relations, *inaccuracy* and *inexactitude* are not synonyms (an *inexactitude* is a kind of *inaccuracy*). In DicoSyn, containing only synonymy relations, *pavillon* (jack) is a synonym of *drapeau* (flag), while *pavillon* can be seen as a particular type of *drapeau*. Nevertheless, results seem acceptable enough for our application.

8 Wisigoth

In order to carry out our enrichment method, we created an architecture called WISIGOTH (WIktionarieS Improvement by Graph-Oriented meTHods) composed of a set of modules depicted in Fig. 4.

carry on?

Fig. 4 The WISIGOTH architecture.

8.1 Computation of Candidates

The first part of the architecture is made of a processing pipeline which builds the graphs introduced in Section 6.1 from a Wiktionary dump. Then it computes the candidate relations by applying the method described in Section 6.2. This processing pipeline can be triggered each time a new dump is released or when a given threshold of edits has been registered.

8.2 Suggestion and Validation of Candidates

The interface we developed to suggest and validate or invalidate new relations is implemented as a Firefox extension. When an user browses the English or French Wiktionary, the interface sends a request to a web service we host, which returns, for each known lexeme, an ordered list of potential synonyms (cf. Fig. 5).

---

[15] The word *chiotte* is a slang version of *cabinet*.

Fig. 5 The WISIGOTH Firefox extension. Example of suggestions for *beautiful.*

Suggestion and Editing: Next to each proposition appears a '+' sign which triggers, when clicked, the automatic addition of the candidate as a synonym to the Wiktionary server. A contributor may want to add a new synonym that has not been suggested, so we provide a free text area. Regardless of our enrichment method, this functionality expands the potential population of contributors. Handling the edition of the wikicode enables all users to become contributors while this opportunity was restricted so far to "wikicode-masters". No cross-validation system, in which a relation would be added only if several contributors validate it, has been designed: To keep close to the wiki principle, we did not add any additional regulation,[16] but as we ease the addition of synonyms, we provide an easy way to remove them too by adding a deletion '-' sign to every synonym occurring in the page.

Notification of editing: Up to now, wiktionaries dumps are released frequently. Nevertheless, we protected against irregular dumps thanks to our interface that notifies the server about synonyms edition. A desynchronization between Wiktionary's current state and our lexical networks could cause irrelevant suggestions. Therefore, a new modeling of synonymy networks and a reprocessing of candidates may be done between two releases.
Storing these notifications will also later give us the opportunity to make further statistical analysis (which synonymy links look problematic, how many users contribute, etc).

Blacklisting: Although we did not rely on a cross-validation system for adding synonyms, we propose a blacklisting system to stop suggesting a candidate judged as irrelevant by several contributors for a given target lexeme. When a candidate is proposed, a contributor may judge it irrelevant and ask for not being proposed again this candidate. This request is stored in the contributor's personal blacklist but the candidate is still proposed to other contributors. When a given threshold of contributors have blacklisted a candidate, this candidate is stored in a global blacklist and is no longer proposed as a synonym of the target lexeme. As a consequence, potentially more relevant candidates may be suggested. The resulting blacklist may be used for later error-analysis of our enrichment method.

An open architecture: Although our system has been first designed to use endogenous data, there is no reason to refrain the use of exogenous data when available. We are including in the graphs data stemming from corpus processing.

If other institutions are willing to join WISIGOTH framework, it is possible for them to provide data to be hosted by our server or to design their own complementary web service that our Firefox extension can request.

---

[16] For some insights into the self-regulation of the Wikiprojects ecosystem, see **?**.

9 Conclusion and Future Work

Observing the lack of satisfying lexical semantic resources, this paper pointed out the problems encountered in their development. Among other difficulties, the evaluation required to validate automatically-built resources is an imperative prerequisite to assess their quality before using them. We have considered the different types of evaluation used in the domain and have shown that only a validation operated by several experts can be reliable: Other evaluations are worth being done, but should be considered as a rough informative guide. Evaluations against gold standards or task-based evaluations of resources introduce some bias hard to overcome, while human-evaluation may lead to low agreement or reasonable agreement that is not always significant. In light of those observations, we proposed a method based on crowdsourcing: Wiktionary, a collaborative dictionary, is used to bootstrap an incomplete synonymy network and we compute new synonymy relations by performing random walks over the network. For the lexemes included in the dictionary, new synonyms can thus be suggested. While the copyleft licence of the online resource solves the problem of availability, relying on crowds of contributors may be a solution to the validation issue. One can object that the contributors' lexical knowledge cannot be guaranteed. However, for languages such as French in which no acceptable resource is available, this solution seems interesting to build a coarse-grained resource. We studied the impact of using several data sources. Methods based on endogenous data makes this approach reproducible for any language and applicable to other lexical networks than Wiktionary. It may help, for example, building WordNets that are under construction, such as the Mandarin Chinese one (**?**). Nevertheless, we do not refrain to use exogenous data when available. Of course, results vary a lot depending on the different data sources used. Some of them are not impressive but are sufficient to be used in our system. The combination of data sources presented can be improved in several ways. Empirical attempts to weight the edges of the combined graphs is tedious and may not lead to an optimum. It would be advisable to rely on machine learning to determine which combination leads to the best result.

It may be surprising that, after having pointed out the bias of using gold standards, we did rely on them for evaluating our system. However, we did not attempt to make a resource evaluation *per se*. We rather used the gold standards to study the impact of data sources on the result and to select the best combination to be used for feeding the suggestions database that the WISIGOTH system requests.

Our purpose was to make a proof of concept. A more relevant evaluation will be possible after one or two years. Indeed, we took the opportunity to study qualitatively and quantitatively the English and French editions of Wiktionary and have shown that they are deficient in terms of synonymy relations. The "real" evaluation will consist on observing whether contributors have used our system and how many synonymy relations has been added with it. In the future, we hope to be in position to present the new curves of the synonymy relations showing an acceleration.

Future work: In the short term, we consider ameliorating our software. When a candidate is proposed as synonym, it may be relevant or not. If not, instead of considering it systematically as noise, it may sometimes correspond to other relations (antonymy, hypernymy, meronymy, etc.) which are hard to differentiate with the automatic methods we use. We plan to add a functionality which adds the candidate in other relation sections (than synonymy section) of the entry. It will permit to enrich the resource and get some insight of what do capture the methods measuring the "semantic relatedness". We have not envisaged so far to study *"non-classical"* relations (**?**), their relevancy outside of what they have been introduced for (lexical cohesion) not being clear to us. Another improvement will be to handle better, with the contributor's help, the word senses sections into which the synonyms are added.

An extension of this work will be the proposition of new translations by leveraging the same kind of graph models and similarity measures. However, the lack of comprehensive gold standards (to our knowledge) will make the evaluation—even rough—difficult and therefore make the development of the new method difficult.

Call for contributions: We would like to foster English and French speakers to test out the Firefox extension we propose. Collected data will be released freely.

Call for collaborations: We have presented in this paper the methods and data we used. We would like to invite anybody willing to join: it can be done by providing candidates to be hosted by our server or by proposing a web service that our Firefox extension could request (cf. Figure 4). Moreover, we are open to collaboration to adapt our system to Wiktionary's other languages or even other lexical resources under construction.

10 Resources

The resources used in and built for this paper are available here:

- `http://redac.univ-tlse2.fr/wisigoth/`
  The WISIGOTH Firefox extension and the structured resources extracted from Wiktionary (English and French)
- `http://redac.univ-tlse2.fr/lexicons/wiktionaryx.html`
  The XML-structured dictionaries extracted from Wiktionary (English and French)
- `http://redac.univ-tlse2.fr/corpora/wikipedia.html`
  Raw-text corpus extracted from the French Wikipedia
- `http://redac.univ-tlse2.fr/applications/vdw.html`
  Syntactic cooccurrents and distributional neighbors computed over the Wikipedia corpus