Semi-automatic Endogenous Enrichment of Collaboratively Constructed Lexical Resources: Piggybacking onto Wiktionary

Franck Sajous[1], Emmanuel Navarro[2], Bruno Gaume[1], Laurent Prévot[3], and Yannick Chudy[1]

[1] CLLE-ERSS, CNRS & Université de Toulouse
[2] IRIT, CNRS & Université de Toulouse
[3] LPL, CNRS & Université de Provence

Abstract. The lack of large-scale, freely available and durable lexical resources, and the consequences for NLP, is widely acknowledged but the attempts to cope with usual bottlenecks preventing their development often result in dead-ends. This article introduces a language-independent, semi-automatic and endogenous method for enriching lexical resources, based on collaborative editing and random walks through existing lexical relationships, and shows how this approach enables us to overcome recurrent impediments. It compares the impact of using different data sources and similarity measures on the task of improving synonymy networks. Finally, it defines an architecture for applying the presented method to Wiktionary and explains how it has been implemented.

Key words: Collaboratively Constructed Lexical Resources, Endogenous Enrichment, Crowdsourcing, Wiktionary, Random Walks.

1 Introduction

While emerging processes of creation and diffusion keep increasing the production of digital documents, the tools to process them still suffer from a lack of acceptable linguistic resources for most languages. *"We desperately need linguistic resources!"* is claimed in [1], after arguing 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, and concluding that a collaborative effort is needed, and that sharing resources is crucial. In this paper, we propose a new method for developing lexical resources which could meet these needs and we apply it to Wiktionary,[4] the free online dictionary. The system we describe automatically computes semantic relations, namely synonyms, to be added or not to a lexical network, after being validated or invalidated by contributors. In Section 2, we take inventory of the usual approaches and point out the impediments that hinder the success of such processes. We then investigate new trends which could help overcome this shortcoming. We outline in Section 3 the key points of our method, based on a

[4] `http://www.wiktionary.org`

semi-automatic endogenous enrichment process. We explain in Section 4 how we compute the candidate relations by random walks over various graphs and using several measures that we evaluate, regarding our specific purpose. We present the architecture built to carry out the whole enrichment/validation system in Section 5 and we describe possible future extensions of out method in Section 6.

2 Lexical Resource Building

2.1 Context

Princeton WordNet [2] is probably the only successful large-scale project among lexical resource building attempts which is widely used. The subsequent projects EuroWordNet [3] and BalkaNet [4] were less ambitious in terms of coverage. Moreover, these resources froze as soon as the projects ended while Princeton WordNet kept on evolving. EuroWordNet's weaknesses have been underlined in [5], and automatic methods to add missing lexical relations have been proposed. Existing resources have been used in [6] to build WOLF, a *free* French WordNet. Pattern-based approaches were first proposed in [7] to harvest semantic relations from corpora and refined in [8] by reducing the need for human supervision. All of the latter three automatic processes would require validation by experts to produce reliable results. However, the cost of this validation work makes it difficult to afford or results in resources that are not freely accessible. The problems of time, cost and availability are increasingly becoming a matter of concern: in corpus-linguistics, an AGILE-like method borrowed from Computer Science has been proposed in [9] to address the problem of simultaneously maximizing corpus size and annotations while minimizing the time and cost involved in the creation of corpora. To tackle the availability issue and build free corpora, a method relying on metadata to automatically detect copylefted web pages is described in [10]. In the domain of lexical resource building, methods relying on *crowdsourcing* may help overcome recurrent bottlenecks.

2.2 Collaboratively Constructed Resources (CCR)

It has been claimed in [11] that the accuracy of Wikipedia comes close to Britanica, who criticized the criteria of the evaluation [12]. A more moderate study [13] has shown in a task measuring the semantic relatedness of words that resources based on the *"wisdom of crowds"* are not superior to resources based on the *"wisdom of linguists"*, but that CCRs are strongly competitive. It has also been demonstrated that "crowds" can outperform linguists in term of coverage.

Collaborative and social approaches to resource building do not rely only on colleagues or students but on random people, who do not share the NLP researchers' interest for linguistic resource building. Therefore, building sophisticated and costly infrastructures that are empty shells waiting to be filled presents the risk of being platforms that no one would visit. 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. However, there are at least two main tracks to follow in order to avoid this pitfall:

Gamers. Some language resource builders have been successful in designing simple web games in which many people come to play just for fun. For instance, the French serious game *"Jeux de Mots"*[5] [14] 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, meronymy, etc.). However, setting up an interesting game for collecting any kind of linguistic information is not easily feasible. For instance, domain-specific resources might be harder to collect this way. Secondly, designing game-play that really works is a difficult task in itself and it is likely that many "game-elicited" resource initiatives will fail because of the game not being fun for random people.

Piggybackers. Only a few collaborative or social infrastructures are really successful. These resources and networks 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 in order to take advantage of their huge amounts of visitors and contributors. Significant steps towards such an architecture have been made in [15, 16]. 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 solid and popular infrastructures requires much less effort and technical skill than setting-up the whole platform (though lots of technical difficulties occur to comply with and plug into these infrastructures).

3 Outline of Proposal for a New Approach

Taking into account the observations made in Section 2 and considering the benefits of using CCRs, we propose a method for enhancing lexical resources that is reasonable in terms of time and cost, based on: (i) piggybacking onto Wiktionary, (ii) computing similarity measures grounded on random walks through the graphs extracted from its lexical networks (Sections 4.2 and 4.3) and (iii) giving an easy way for users to validate the candidate relations that we suggest.

3.1 Wiktionary

Wiktionary is a free multilingual collaborative dictionary including definitions, semantic relations and translations (a detailed presentation can be found in [15, 16]). Its intrinsic features fulfill some of our needs: it is publicly available, its growth is fast and continuous and, as its content is based on crowdsourcing, the *"reasonable cost"* constraint turns euphemistic. However, what is the quality of resources constructed by "naive speakers" as compared to those built by skilled professional lexicographers? A recent study [17] evaluated three German resources designed in different manners: expert-built (GermaNet), semi-controlled (OpenThesaurus) and collaboratively edited (German Wiktionary). This comparison demonstrated that all resources have a similar topology[6] and lexical

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

[6] extracted graphs are small worlds with a heavy-tailed degree distribution: see [18].

coverage, but different density of semantic relations: for instance, Wiktionary has fewer hypernyms/hyponyms than GermanNet, but clearly outperforms both other resources in term of antonymy relations. Table 1 gives the number of common nouns, verbs, adjectives and (undirected) synonymy and translation links for the French and English Wiktionaries in 2008 and 2010. These figures relate to all lexemes found—conversely, in [16], only the lexemes connected by synonymy links have been counted. Translation and synonymy links have been counted after the graphs have been symmetrized (i.e. two-way links are counted once).

Table 1. Growth of French and English Wiktionaries from year 2008 to 2010.

		2008			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)
	Synonymy links	9 670	1 793	2 522	17 054 (×1.8)	3 158 (×1.8)	4 111 (×1.6)
	Translation links	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)
	Synonymy links	12 271	3 621	4 483	28 193 (×2.3)	8 602 (×2.4)	9 574 (×2.1)
	Translation links	172 158	37 405	34 338	277 453 (×1.6)	70 271 (×1.9)	54 789 (×1.6)

As we can see, the number of lexemes has seen a growth that makes Wiktionary, for these languages, comparable to commercial printed dictionaries in term of lexical coverage: the French *"Petit Robert"* includes 60 000 entries and the *"Longman Dictionary of Contemporary English"* features 50 000 entries. Moreover, all the resources that capture some aspect of linguistic knowledge can prove to be useful and interesting. So, traditional resources and collaborative resources should both continue to be developed, especially since, as mentioned by [19], their content does not overlap too much.
Regarding semantic relations, we have shown the sparseness of the synonymy networks extracted from Wiktionary in 2008 [16]. Synonymy relations grew at slower rate than lexeme coverage, which makes the 2010 graphs even more sparse. To help fill this gap, we present below an endogenous enrichment method.

3.2 Endogenous Enrichment

Our aim is to be able to propose, for an existing semantic lexical network, new relations that are potentially missing. To propose new pairs of words which may be synonymous, we compute a similarity measure between any two nodes (lexemes) of the network by applying random walks through already existing lexical relations. Details of the different data sources, graph modeling and measures we use are given in Section 4.
As the potential new synonyms we compute are to be validated by contributors, and not automatically added to the initial resource, our purpose is: (i) to suggest candidates for the greatest number of lexemes and (ii) for a given lexeme, to propose a finite list of candidates including at least *"some"* relevant ones.
In our case, it is better to propose no candidate at all than irrelevant ones, and the system is not meant to suggest *all* relevant candidates: first, because a contributor won't check an endless list and secondly, our method is an iterative computation-suggestion-validation cycle. Thus, if a relevant candidate is not

initially proposed, it may be the next in the list of suggestions, which may be shifted when a suggested candidate is chosen. So, as the relations added to the network will change its structure, and as the computation of candidates will be reprocessed regularly (after the release of a new dump in the case of Wiktionary), this relevant candidate may be proposed after some iterations. Thus, recall will increase with successive iterations and we focus therefore more on precision.

3.3 Validation

The candidates that we compute are suggested to the contributors via an interface described in Section 5. If a contributor validates a suggestion, the relation is added to Wiktionary. 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,[7] but as we ease the addition of synonyms, we fairly give an easy way to remove them too.

4 Similarity Elicitation

This section presents the methods used to compute, from existing lexical networks, new synonymy relations to be added. We rely on different kinds of data and similarity measures and compare the results obtained by evaluating them against expert-built gold standards.

4.1 Data

Networks have been extracted from English and French Wiktionaries for nouns, verbs and adjectives, thus splitting the global structure of the dictionaries into mono-part of speech subparts. Given a language version of Wiktionary, we consider only the article sections dedicated to entries in the language of interest, e.g. the English lexemes of the English Wiktionary. From these sections, we extract the existing synonymy and translation links, as well as the glosses.

4.2 Bipartite Graphs Model

In order to homogenize and simplify the description of the experiments, each type of data we used will be modelled as a *undirected bipartite graph* $G = (V \cup V', E)$ where the set of vertices V will always denote the lexemes of the language and part of speech of interest, whereas another set of vertices V' will vary depending on the sources of data. The set of edges E is such that $E \subseteq (V \times V') \cup (V' \times V)$ and models the relations between the lexemes of V and of V'.

• **Translation graph** $\mathbf{G_{Wt}} = (\mathbf{V} \cup \mathbf{V_{Wt}}, \mathbf{E_{Wt}})$. 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 undirected edge between $v \in V$ and $t \in V_{Wt}$ if t is found as a translation of v.[8]

[7] For some insights into the autoregulation of the Wikiprojects ecosystem, see [20].

[8] As we parse only the dump of the language of interest, we find the *oriented* link $v \rightarrow t$ (t as a translation of v in v's article) and symmetrize it into $v \leftrightarrow t$. Having a more subtle model (with oriented edges) requires parsing all dumps of all languages.

• **Synonymy graph $\mathbf{G_{Ws}} = (\mathbf{V} \cup \mathbf{V_{Ws}}, \mathbf{E_{Ws}})$**. Here, $V' = V_{Ws}$ is a copy of V. There is an undirected edge between $v \in V$ and $u \in V_{Ws}$ either if $v = u$ or if u (or v) is indicated as a synonym in v (or u) entry. This bipartite graph model of the synonymy network may look unusual, however: (i) it permits us 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 $\mathbf{G_{Wg}} = (\mathbf{V} \cup \mathbf{V_{Wg}}, \mathbf{E_{Wg}})$**. Here, $V' = V_{Wg}$ corresponds to the set of all lemmatized lexemes found in the glosses of all entries. There is an undirected 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, its glosses have been concatenated, lemmatized and tagged with Treetagger,[9] and stopwords have been removed.

• **Syns+Trans graph $\mathbf{G_{Ws+t}} = (\mathbf{V} \cup \mathbf{V_{Ws}} \cup \mathbf{V_{wt}}, \mathbf{E_{Wst}} = \mathbf{E_{Ws}} \cup \mathbf{E_{Wt}})$**. Here, $V' = V_{Ws} \cup V_{wt}$ includes a copy of the set of lexemes V and their translations. There is an undirected edge between $v \in V$ and $v' \in V_{Ws} \cup V_{Wt}$ if v' is either a synonym or a translation of v.

4.3 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 intent is to propose as candidates the pairs with the highest scores (which are not already known as synonyms in Wiktionary). We test various similarity measures, all based on—short—fixed length random walks. Such approaches for measuring the "topological resemblance" in graphs are introduced in [18, 21]. This kind of methods is applied to lexical networks in [22] to compute semantic relatedness. We consider a walker wandering at random in the *undirected bipartite graph* $G = (V \cup V', E)$, 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:

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

where $d(i)$ is the degree of incidence (number of neighbours) of vertex i. 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 afer t steps, starting from v. This is the first measure[10] (called *simple*) we use ; other measures are based on this one:

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

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

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

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

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

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

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

"cos" and "dot" are respectively the classical cosine and scalar product. "ZKL_γ" is a variant of the Kullback-Leibler divergence introduced in [22].
Let $C(v, G, t, \mathrm{sim})$ 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:

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

where E_{Ws} is the set of existing synonymy links in Wiktionary. The experiments below consist in evaluating the relevancy of $C(v)$ when G and sim vary. $t = 2$ will remain constant.[11]

4.4 Evaluation method

In view of our application (cf. Section 5.2) and given the criteria defined in Section 3.2, for each lexeme, we consider that a *suggested list* of candidates is *acceptable* if it includes at least one relevant candidate. Indeed, a 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 candidates.

Let $G_{GS} = (V_{GS}, E_{FS})$ 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 "known" by the gold standard (i.e. $v \in V_{GS}$). Indeed, if a lexeme $v \in V$ does not belong to the gold standard (i.e. $v \notin V \cap V_{GS}$), we consider that it is a lexical coverage issue, so one cannot deem whether a relation (v, c) is correct or not.[12] For the same reason, for each lexeme v, we remove from $C(v)$ the candidates 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 $\Gamma_k(v)$ the "evaluable" suggested list of candidates:

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

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

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

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

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

We define the set N_k of lexemes having at least one candidate being proposed and the set N_k^+ of lexemes for which at least one *correct* candidate is proposed:

$$N_k = \left\{v \in V \cap V_{GS} / \Gamma_k(v) \neq \varnothing\right\}, N_k^+ = \left\{v \in V \cap V_{GS} / \Gamma_k^+(v) \neq \varnothing\right\} \tag{10}$$

To compare the efficiency of different data sources used to compute the candidates, we measure P_k, the ratio between the *acceptable* suggested lists and the lexemes for which suggestions are done, and R_k, the ratio between the number of suggested lists and the number of evaluable target lexemes:

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

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.

4.5 Results

Gold Standards: We used Princeton WordNet to evaluate the candidates for English and DicoSyn[13] for French. The extraction of the synonymy networks from these resources reproduces what has been done in [16].

Similarity measures: Applying the different similarity measures presented in Section 4.3 shows that all give pretty similar results. As an example, the results obtained for the intersection of the gold standards and the English and French Wiktionaries' nouns and verbs are reported in Table 2. The *simple* measure being as efficient as the others and having far less complexity, further experiments have therefore been done using this measure.

Table 2. 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	41.4	32.4	58.6	47.3	51.4	37.8	**78.7**	58.3	51.9	39.0	**74.6**	**55.3**
avg	42.5	33.5	58.2	46.8	50.5	38.0	**78.7**	58.3	51.1	39.3	74.0	55.1
cos	**43.4**	**34.6**	**60.2**	47.9	51.8	38.5	78.3	58.6	51.3	39.4	73.1	54.2
dot	42.0	34.0	59.7	46.7	**52.3**	**38.7**	78.2	58.7	**52.4**	39.7	73.6	54.8
ZKL_{10}	43.2	34.0	60.1	**48.2**	51.8	38.6	**78.7**	**58.8**	51.9	**39.8**	74.0	54.5

Data sources: As we can see in Table 3, better results are obtained for French than for English. This can be partly explained by the slightly lower density of the English networks (cf. Table 1) but is mainly due to the difference between the gold standards used: networks extracted from WordNet are more sparse than

[13] Dicosyn is a compilation of synonym relations extracted from seven dictionaries produced at ATILF and corrected at CRISCO units.

the ones extracted from Dicosyn (see [16]). Morever, Table 4 shows that some candidates rejected by the gold standards do not look unreasonable, which makes it hard to draw definitive conclusions. Nevertheless, despite a—potentially—severe evaluation, results look acceptable enough in view of our application.
The translations graph provides better precision than synonymy graphs. This result was expected, as in Wiktionary, lexemes have more translation links than synonyms. Moreover, translations are often distributed over several languages, which is more reliable than having a lot of translations into a given language.
The glosses graph's worse precision and higher recall was expected too: almost all lexemes have glosses, but information is less specific, and we did not try any tricky edge weighting. Combining synonyms and translations enables a better recall than with separated graphs and a similar precision for English. For French, it leads to a loss of precision compared to the "translations only" graph.

Table 3. Impact of different data sources on the *simple* similarity measure.

					Synonyms		**Translations**		**Syn.+Trans.**		**Glosses**	
		$\lvert V\rvert$	$\lvert V_{GS}\rvert$	$\lvert V \cap V_{GS}\rvert$	P_5	R_5	P_5	R_5	P_5	R_5	P_5	R_5
EN	Adj.	48930	21479	13742	46.3	24.9	53.5	23.4	53.7	34.6	26.1	98.3
	Nouns	196790	117798	43236	32.4	17.1	37.8	24.9	39.0	32.4	14.9	98.9
	Verbs	67649	11529	8890	41.4	33.2	51.4	43.5	51.9	53.8	27.0	99.9
FR	Adj.	41725	9452	3958	61.2	24.9	76.1	19.8	69.6	34.2	32.2	96.1
	Nouns	106068	29372	16084	47.3	23.2	58.3	22.2	55.3	35.4	20.7	99.4
	Verbs	17782	9147	4037	58.6	22.3	78.7	36.8	74.6	45.8	41.1	99.4

Table 4. Example of propositions for nouns evaluated against gold standards (GS).

	in GS	**Propositions**
EN	Yes	*<imprisonment: captivity>, <harmony: peace>, <filth: dirt>, <antipasto: starter>, <load: burden>, <possessive: genitive>*
	No	*<rebirth: renewal>, <fool: idiot, dummy>, <cheating: fraud>, <bypass: circumvention>, <dissimilarity: variance>, <pro: benefit>*
FR	Yes	*<ouvrage, travail>, <renom: gloire>, <emploi: fonction>, <drapeau: pavillon>, <rythme: cadence>, <roulotte: caravane>, <chinois: tamis>*
	No	*<drogue: psychotrope>, <fantassin: bidasse>, <force: poigne>, <salade: bobard>, <W.C.: chiotte>, <us: tradition>, <bisque: soupe>*

5 Implementation: the WISIGOTH Architecture

In order to carry out our enrichment method, we designed an architecture called WISIGOTH[14] composed of a set of modules depicted in Fig. 1.

5.1 Computation of Candidates

The first part of the architecture is made of a processing pipeline which, from a Wiktionary dump,[15] builds the graphs introduced in Section 4.2 and computes the candidate relations by applying the method described in Section 4.3. This processing pipeline can be triggered each time a new dump is released or when a given threshold of edits has been registered.

[14] WIktionarieS Improvement by Graph-Oriented meTHods.
[15] Wiktionaries' dumps are available at: `http://download.wikipedia.org/`

5.2 Suggestion and Validation of Candidates

The interface we developed to suggest and validate or invalidate new relations materializes as a Firefox extension. Once installed, when a user browses the English or French Wiktionary, the interface sends a request to the candidates service which returns, for each known lexeme, a list of potential synonyms.

Suggestion and Editing: Next to each proposition appears a '+' sign which triggers the automatic addition of the candidate as a synonym to the Wiktionary server. As a contributor may want to add a synonym that has not yet been suggested, we provide a free text area too. Regardless of our enrichment method, it enlarges the potential population of contributors, not restricting it to "wikicode-masters" acquainted with the underlying syntax. As explained is Section 3.3, a '−' sign is added to every synonym occuring in the page, which handles the deletion of this synonym.

Notification of editing: Thus far, wiktionaries dumps are released frequently. Nevertheless, to protect against irregular dumps which could result in a desynchronization between Wiktionary's current state and the lexical networks we extracted from it—and therefore, cause irrelevant suggestions—, the interface notifies our server the editing of synonyms. Thus, a remodelling 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 analyse which synonymy links look problematic (e.g. a series of additions and deletions) and how contributors behave.

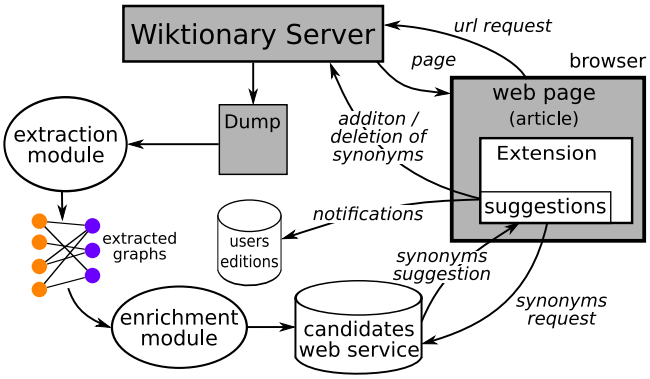

Fig. 1. The WISIGOTH architecture.

6 Conclusion and Future Work

This paper has pointed out the problems usually encountered in the development of lexical resources. It has shown how CCRs help overcome these difficulties and, among them, how we can take advantage of Wiktionary's infrastructure and content. Nevertheless, "crowds" are more prone to add new words than to provide semantic relations. To encourage them, we have designed a tool to assist

collaborative editing by suggesting new synonyms to be validated. We took the opportunity to compare the impacts of using different data sources and similarity measures. The choice of the measure does not much affect the results whereas combining data sources permits us to gain precision or recall, depending on the language. Adding glosses to the *"Syn+Trad"* graph presented and working on the weighting of the graphs' edges should bring even better results.

Grounded on the topology of the graphs extracted from the lexical networks, this system is language-independent and, moreover, may be applied to other resources than Wiktionary, contrary to methods like [23] which exploit the structure of hyperlinks between pages and are therefore bound to this resource. It may help, for example, building WordNets that are still under construction, as the Chinese one [24]. Moreover, not relying on other external resources makes this method endogenous and may be applicable to enhance lexical resources for under-resourced languages. When external resources are available, for example stemming from distributional analysis over large corpora, an exogenous enrichment module can be coupled to our system and feed our edition interface.

A short-term extension of this work will be the proposition of new translations by leveraging the same kind of graph model and similarity measures. Linguistic observations should be done to characterize what other kinds of semantic relation (than synonymy) is captured by automatically computed relatedness.

Although we did not rely on a cross-validation system for adding synonyms, we think it could be useful to add a blacklist system to stop proposing a candidate judged as irrelevant by several contributors for a given target lexeme.

An interesting study would be the evaluation of the results of the endogenous enrichment process at different stages of Wiktionary's growth. This can be done by rebuilding the various past states of the lexical networks using the "historical dump" containing all articles revisions. Such a study may show when it is appropriate to apply our method: when we have enough material to start suggesting new relations and when no more relevant relation is to be proposed and should be stopped.

Resources: the Firefox extension presented in this paper and the structured data extracted from Wiktionary's dumps are publicly available at:
`http://redac.univ-tlse2.fr/wisigoth/`

References

1. Sekine, S.: We desperately need linguistic resources! –based on the users' point of view. FLaReNet Forum 2010. Barcelona, Spain (2010)
2. Fellbaum, C., ed.: WordNet: An Electronic Lexical Database. MIT Press (1998)
3. Vossen, P., ed.: EuroWordNet: a Multilingual Database with Lexical Semantic Networks. Kluwer Academic Publishers, Norwell, MA, USA (1998)
4. Tufis, D.: Balkanet Design and Development of a Multilingual Balkan Wordnet. Romanian Journal of Information Science and Technology **7** (2000)
5. Jacquin, C., Desmontils, E., Monceaux, L.: French EuroWordNet Lexical Database Improvements. In: Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING). (2007) 12–22

6. Sagot, B., Fišer, D.: Building a Free French Wordnet from Multilingual Resources. In: Proceedings of OntoLex 2008, Marrakech (2008)
7. Hearst, M.A.: Automatic Acquisition of Hyponyms from Large Text Corpora. In: Proceedings of the 14th International Conference on Computational Linguistics (COLING), Nantes (1992) 539–545
8. Pantel, P., Pennacchiotti, M.: Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. In: Proceedings of the International Conference on Computational Linguistics, Sydney, ACL Press (2006) 113–120
9. Voormann, H., Gut, U.: Agile Corpus Creation. Corpus Linguistics and Lingustic Theory **4** (2008) 235–251
10. Brunello, M.: The Creation of Free Linguistic Corpora from the Web. In: Proceedings of WAC5: 5th Workshop on Web As Corpus, San Sebastian (2009) 37–44
11. Giles, J.: Internet Encyclopaedias Go Head to Head. Nature **438** (2005) 900–901
12. Encyclopaedia Britannica: Fatally Flawed: Refuting the Recent Study on Encyclopedic Accuracy by the Journal Nature (2006)
13. Zesch, T., Gurevych, I.: Wisdom of Crowds versus Wisdom of Linguists – Measuring the Semantic Relatedness of Words. Journal of Natural Language Engineering. **16** (2010) 25–59
14. Lafourcade, M.: Making People Play for Lexical Acquisition with the JeuxDeMots prototype. In: SNLP'07: 7th International Symposium on Natural Language Processing, Pattaya, Thailand (2007)
15. Zesch, T., Müller, C., Gurevych, I.: Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In: Proceedings of the Conference on Language Resources and Evaluation (LREC), Marrakech (2008)
16. Navarro, E., Sajous, F., Gaume, B., Prévot, L., Hsieh, S., Kuo, I., Magistry, P., Huang, C.R.: Wiktionary and NLP: Improving Synonymy Networks. In: Proceedings of the ACL-IJCNLP Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources, Singapore (2009) 19–27
17. Meyer, C.M., Gurevych, I.: Worth its Weight in Gold or Yet Another Resource – A Comparative Study of Wiktionary, OpenThesaurus and GermaNet. In: Proceedings of the 11^{th} International Conference on Intelligent Text Processing and Computational Linguistics. Volume 6008 of LNCS. Springer (2010) 38–49
18. Gaume, B., Venant, F., Victorri, B.: Hierarchy in Lexical Organization of Natural Language. In Pumain, D., ed.: Hierarchy in Natural and Social Sciences. Methodos series. Kluwer Academic Publishers (2005) 121–143
19. Zesch, T.: What's the Difference? Comparing Expert-Built and Collaboratively-Built Lexical Semantic Resources. FLaReNet Forum 2010. Barcelona, Spain (2010)
20. Forte, A., Bruckman, A.: Scaling Consensus: Increasing Decentralization in Wikipedia Governance. In: Proceedings of the 41st Hawaii International Conference on System Sciences, Washington DC, IEEE Computer Society (2008) 157
21. Gaume, B., Mathieu, F.: PageRank Induced Topology for Real-World Networks. Complex Systems (2008)
22. Hughes, T., Ramage, D.: Lexical Semantic Relatedness with Random Graph Walks. In: Proceedings of EMNLP-CoNLL. (2007) 581–589
23. Weale, T., Brew, C., Fosler-Lussier, E.: Using the Wiktionary Graph Structure for Synonym Detection. In: Proceedings of the ACL-IJCNLP Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources, Singapore (2009) 28–31
24. Huang, C.R., Chen, C.L., Weng, C.X., Lee, H.P., Chen, Y.X., Chen, K.J.: The Sinica Sense Management System: Design and Implementation. Computational Linguistics and Chinese Language Processing **10** (2005) 417–430