NLP lexicons: innovative constructions and usages for machines and humans

Nuria Gala, Mathieu Lafourcade

NLP lexicons: innovative constructions and usages for machines and humans

Nuria Gala[(1)], Mathieu Lafourcade[(2)]
[(1)]LIF-CNRS, 163 av. de Luminy case 901, 13288 Marseille Cedex 9, France
[(2)]LIRMM-CNRS, 161 rue Ada, 34392 Montpellier Cedex 5, France
Email: nuria.gala@lif.univ-mrs.fr, mathieu.lafourcade@lirmm.fr

Abstract

Lexical resources have undergone significant changes with the generalized use of computers and the advent of the Internet. However, while such changes stand for revolutions when it comes to compare machine-readable dictionaries to their paper 'ancestors', machine-readable dictionaries, compiled for human readers, still have serious limitations. Natural language processing lexicons, initially developed for NLP applications, have shed light on some of such shortcomings. In this presentation, we will attempt to bring new elements relatively to NLP approaches aiming to develop present and tomorrow's lexical resources, in particular, using morphological and semantic information to better access lexical items. A special focus will be given on the semantic and on the multilingual side. Our argument is that nowadays lexical resources 1) should be useful both for men and machines, 2) can be constructed in alternative ways from classical lexicographic work, and 3) provide novel accesses and usages that are feasible only in the context of computer and user networks. Such points will be highlighted by means of two resources under development: *LexRom,* as an example of morphological form-based multilingual access, and the lexical network of *JeuxDeMots,* as an illustration of associative and semantic access.

Keywords: NLP lexica; crowd sourcing; semi-supervised learning; morphological and semantic content; multilingual and semantic access

1. Introduction

Since the introduction of computers, lexical resources have undergone significant changes in terms of lexicographical practices and user access to words. For more than thirty years, the growing contribution of computers to lexicography has transformed the way to create and enrich lexical resources[1]. Yet the impact of using machines into the lexicographic field has already been discussed in the literature by leading contributors (Atkins & Zampolli, 1994; Grefenstette, 1998; Rundell, 2002, among others). However, the subject still remains on the table, mainly because the achievements in terms of the resources themselves are far from being as satisfactory as the electronic media could entail.

While lexicographical practices have significantly evolved due to the access to large amounts of data and the use of highly-skilled and linguistically-aware editors (Rundell, 2002), machine-readable dictionaries (MRDs) still stand for electronic versions of their paper 'ancestors'. Doubtless, the use of large amounts of data allowed the incorporation of new information into the dictionaries: statistical – frequencies – (Kilgarriff, 1997), collocational behaviors, and even much complex patterns – syntactic patterns – gathered by corpus query tools (Jakubíček et al., 2010). Furthermore, the electronic media involved the combination of multimedia lexicographic material like sounds (pronunciations), images, videos (sign languages), etc. which might be of help in particular contexts and for specific users: foreign learners (i.e. Merriam Webster[2]), deaf people (i.e. Arasaac[3], Tegnspro[4]), among others. Nevertheless, it should be noted, as Grefenstette (1998) did, that the lexicon represented in the dictionaries is still seen as two-dimensional: "a list of lists" (Heid, 2009), that is to say, a list of words with their associated explanations, be them linguistic, statistical or multimedia.

As far as the access to the lexicographic information is concerned, the major revolution is the fact that the user has a variety of research criteria going from a target key word to more complex search patterns (a specific domain, a grammatical category, etc.). Frequently, s/he may even choose among several possibilities. Yet, in spite of such functionalities, alphabetical lists generally remain, as if the user was unable to get rid of traditional habits.

More recently, online interactive dictionaries appear to be real platforms giving access to several interconnected lexical resources. One example is the *Nuevo Tesoro Lexicográfico de la Lengua Española*[5] (NTLLE), a resource from the Real Academia Española grouping about 70 dictionaries resulting from five centuries of institutional Spanish lexicography. Another example might be *Wordnik*[6], a resource giving access to a variety of lexical resources (dictionaries, corpora, thesauri, etc.) and thus going beyond the user expectations with "as much information as possible" about a word.

[1] The *Trésor de la Langue Française* (1971-1994) innovated French lexicography with the use of computer indexing of a wide corpus of texts (Frantext). The *Collins COBUILD* (1987) was the first dictionary where electronic corpora was used, thus providing primary English data source (7 million word).

[2] http://www.merriam-webster.com
[3] http://www.catedu.es/arasaac
[4] http://tegnsprog.dk
[5] http://buscon.rae.es/ntlle/SrvltGUILoginNtlle
[6] http://www.wordnik.com

Despite such significant progress resulting from computer means (no need to mention increasing storage capabilities, nor reduced response time for a query), MRDs, enriched with hyperlinks but still compiled for human readers, remain two-dimensional repositories and have serious limitations (Fellbaum & Miller 2003), namely on the access to the semantic content otherwise than through words (or their pronunciations / grammatical information) and also on the granularity of the information (they do not include information that they assume the user knows). Additional shortcomings can be raised about the size of the resources in terms of language coverage as well as on the cross-lingual equivalencies.

In this paper, we will attempt to shed light on some of such shortcomings by bringing new elements relatively to NLP approaches aiming to develop present and tomorrow's lexical resources. In particular, a special focus will be given on the semantic and on the multilingual side, using morphological and lexical information to better access lexical objects. The paper is structured as follows. First, MRDs and NLP lexicons are compared. Second, alternatives to human (lexicographic) constructions are dealt with. The following sections are devoted to two lexical resources to illustrate the points already highlighted on the previous sections: accessing to words by their form in a multilingual context (*LexRom*) and through lexical functions (*JeuxDeMots*). The paper concludes by a look at open questions and current developments.

2. Natural Language Processing (NLP) lexicons: from lists to networks

MRDs have been used as a source to collect lexical knowledge for a variety of natural language processing (NLP) applications. Yet considerable research has been done for more than thirty years on automatic extraction of structured knowledge from MRDs: lexical relations, semantic information, taxonomies, etc. However, as Ide & Véronis already pointed out, the results of MRD research for NLP failed to live up to early expectations: "encouraging line of research" (Véronis & Ide, 1990) but "the information they [MRDs] contain is both too inconsistent and incomplete to provide a ready-made source of comprehensive lexical knowledge" (Ide & Véronis, 1994).

At the same time, as they are valuable sources of linguistic information, the NLP community has been actively involved in the creation of a wide range of lexical resources (from computational lexicons –lexical databases– to annotated corpora). While initially created for machine applications, such resources may also be of interest for humans through appropriate interfaces (i.e. language learning, speech therapies, linguistic studies, etc.). NLP researchers have thus been developing computational lexicons leading to significant advances not only on construction methods and techniques (see section 3) but also on the resources themselves: lexical databases are conceived bearing in mind their primary purpose, that is NLP applications, which entails automating –as much as possible– the process of linguistic data analysis with robust technologies. As a result, the information is *structured*, *explicit* and *multi-dimensional*, moving "from lists to networks of lexical objects" (Heid, 2009). To put it in other words, a variety of information is scattered throughout different levels and the user browses to specific contents depending on his/her needs. Lexical resources are thus increasingly *dynamic* as the information is interconnected and available by different means (see sections 4 and 5).

In recent years, a significant number of NLP lexical resources have been developed in a large-scale perspective. Series of projects (EAGLES, MULTEXT, etc.) have converged to standards and models to provide a common framework for their construction, maintenance and extension, i.e. the *Lexical Markup Framework* (LMF) (Francopoulo et al., 2006). These models address linguistic representation and encoding guidelines at different layers (morphology, syntactic behaviors, semantic organization, etc.). Significant projects have come to life, *WordNet* (Fellbaum, 1990) being one of the most outstanding.

3. NLP methods for development, enrichment and evaluation of lexical resources

Due to the nature of language, large-scale lexicon development poses difficult challenges (Calzolari et al., 1999). As manual development is very costly and time consuming, automatic and collaborative building of computational lexicons are real alternatives.

3.1 Automatic acquisition of linguistic knowledge

The cost of manual elaboration and enrichment of resources is generally put forward as a major inconvenience. Within the context of lexical resources and NLP tools development, a response to such uneasiness is the automatic acquisition of linguistic knowledge. Over the last twenty years, a number of unsupervised and (semi-)supervised approaches have become a real alternative yielding to encouraging results at different linguistic layers: morphology (Clément et al., 2004), syntax (Briscoe & Carrol, 1997), semantics (Navigli et al. 2003). The overall idea is to induce linguistic knowledge from available data. Depending on the characteristics of the underlying data (raw or annotated corpora, lexical databases, MRDs, etc.), the target resource would be developed more or less straightforward. In any case, manual evaluation would be necessary, but once again, at different degrees depending on one hand, on the underlying data and, on the other hand, on the aimed granularity of the linguistic description. The more explicit the underlying data, the more explicit the target resource, though more difficult its development.

Due to the availability of large amounts of corpora, statistical models have been playing a major role within the NLP community. Unsupervised techniques do not presuppose explicit linguistic knowledge (annotations): they allow the acquisition of linguistic information from raw corpora. If the results are below other approaches that use annotated data, the major advantage is the availability of unlabeled data (i.e the Web). Very often, such methods are used as a first step for preprocessing raw corpora or to incrementally improve the models .

Semi-supervised approaches exploit some kind of information already encoded or annotated on corpora (i.e. part-of-speech tags). Such methods yield to better results than unsupervised ones because the underlying data allows to induce better linguistic information. In many cases, automatic acquisition of linguistic knowledge for lexical development is based on combining both unsupervised and semi-supervised approaches (see Section 4).

Finally, an alternative to the use of corpora is the use of existing lexical resources. While a number of projects have come to light by using MRDs – with mixed results already mentioned on the previous section – , the use of existing computational lexicons is an interesting option as linguistic knowledge is made explicit. Such line of research is currently widespread, though the resources are not always easily available.

3.2 Collaborative approaches

Collaborative resources, i.e. *Papillon* (Boitet et al., 2002), are based on the principle of sharing contributions, that is, anyone collaborates to enrich the database according to his/her possibilities. The insights of this philosophy are interesting but the results are sometimes disappointing as enriching a resource may become tedious very quickly, and in practice people tend not to participate. Hence, it is hard to get the expected volume of contribution (Cristea et al., 2008).

Over the last decade, the web has led to collaborative projects (*wikis*) based on the participation of volunteers under the supervision of an administrator. Significant projects as regards to lexical semantic resources can be mentionned: *OntoWiki*[7] and *Anawiki*[8] (Poesio et al., 2008) among others. However, if such approaches are appropriate for resources of reasonable size and very good quality (gold standards), they are less suitable for large-scale development (Fort et al., 2010).

One way to avoid such a drawback may be crowd-sourcing through gaming, i.e. games with a purpose (GWAP). In such approaches, volunteers are motivated throughout competition (see Section 5).

Lastly, a new trend has emerged which consist on on-line microworking (a task is cut into small pieces and their execution is paid for). *Mechanical Turk* is one such systems: since its introduction in 2005 it has been increasingly being used for building and validating NLP resources at very low cost (Fort et al., 2010), e.g. transcription, word sense disambiguation, compound relations annotation, categorization, etc. However, a number of drawbacks are being brought to light, namely the small number of trained annotators and thus the annotation quality of the resources produced that way: "if a microworking system is considered desirable by the ACL and ISCA communities, then we also suggest that they explore the creation and use of a linguistically specialized special-purpose microworking alternative to MTurk that both ensures linguistic quality and holds itself to the highest ethical standards of employer/employee relationships" (Fort et al., 2010).

As a first conclusion, lexical resources can be constructed in alternative ways from classical lexicographic work and may be used both for men and machines. Novel accesses and usages may be thus provided, feasible only in the context of computer and user networks. *LexRom* and *JeuxDeMots* appear to be obvious examples.

4. LexRom

LexRom (Gala, 2011) is a project of a multilingual lexicon for Romance languages based on family clusters, providing morphological and semantic information on word families crosslingually. The project aims to be of help in contrastive linguistic research as well as in different NLP and human applications, going from crosslingual information retrieval to interlingual language learning. Spanish and Catalan families have been automatically acquired from corpora and monolingual lexicons, from an initial list of manually encoded words from the French morphological resource *Polymots* (Gala et al., 2010)[9].

To our knowledge, attempts to build multilingual lexical resources have mainly focused on semantic relations between concepts among different languages, i.e. *EuroWordNet* (Vossen, 1998). Other interesting proposals merge lexical and encyclopedic knowledge automatically extracted from WordNet and Wikipedia, i.e. *Babelnet* (Navigli and Ponzetto, 2010). As for morphology, the reference multilingual database is *Celex* (Baayen et al., 1995), yet it has been created as three separated lexicons for English, Dutch and German and thus no interlingual links are available.

4.1 Word-forms and semantic cues

The notion underlying *LexRom* is that of morpho-phonological families. A morpho-phonological family

[7] http://ontowiki.eu/
[8] http://anawiki.essex.ac.uk/

[9] http://polymots.lif.univ-mrs.fr

groups together lexical units sharing phonological, morphological and semantic features. Such a family is usually built around a common stem. For example, in French, the stem 'olive' will induce the family made of lexical units such as 'olivaison' (olive harvesting), 'oliveraie' (olive grove), 'olivier' (olive tree)[10], etc. (see first line on Table 1). For each lexical entry in a family, the following types of information is displayed:

- **morphological structure**: i.e. for 'olivier', base-form *oliv-*, affixes *-i* and *-er* ;
- eventual **phonological alternations**: i.e. 'fleur/flor-' is the stem for words such as 'fleur' (flower), 'fleurir' (bloom) and also 'floraison' (flowering); 'croc/croch-' is the stem for 'croc' (hook) and 'accrocher' (hang);
- **semantic cues**: semantic units associated to the target entry (i.e. for 'olive tree': tree, olive, Jerusalem, etc.). The semantic cues enable to distinguish semantic clusters within a morphological family: words associated to the same idea (see Table 1 for Catalan and Spanish: unlike French, in these languages "oil" and "olive" are two clusters within the same morphological family).

In addition to the linguistic information on lexical entries, for each family, it is possible to see the number of derived items, the number of semantic clusters as well as an indication about how productive the stem might be (low, middle, high).

4.2 Bunches of words cross-lingually

LexRom displays word-families across languages. We thus consider the organization of the lexicon of a language as a set of "bunches of words" sharing a common stem and conceptual fragments. Our hypothesis is that such organization may be found across languages, particularly across closely-related ones. The data obtained will enable to give evidence on (mis)matches in terms of family sizes, lexical holes, equivalent clusters and specific phenomena concerning languages in contrast.

FR	*olive, olivade, olivaire, olivaie, olivaison, olivâtre, oliver, oliveraie, olivette, oliveur, olivier, olivine*
CA	*oli, oliada, oliaire, oliar, oliós, oliva, olivaci, olivaire, olivar, olivarda, olivarer, olivari, olivater, oliveda, olivella, olivellenc, oliver, olivera, oliverar, olivereda, oliverer...*
ES	*aceite, aceitadora, aceitar, aceitera, aceitero, aceitillo, aceitoso, desaceitar, aceituna, aceitunado, aceitunero, aceitunillo, aceituno*

Table 1: "Olive" family for French, Spanish and Catalan

As for lexical productivity, significant differences can come to light by observing the data in *LexRom*[11]. Table 1 shows as example the word 'olive': in Catalan and Spanish such form produces derived words for two different semantic clusters, 'olive' and 'oil' (with two different stems, *oli-* in Catalan and *aceit-* in Spanish). However, the corresponding stem in French (*olive*) only produces derived forms of the 'olive' family (the 'oil' family uses another stem, *huile,* and thus creates another word family). Similarly, *aguja* ('needle') and *agujero* ('hole') are part of the same word family in Spanish although in other similar Romance languages the semantic meaning of the latter is conveyed by means of different word forms: *buco* (Italian), *forat* (Catalan), *trou* (French), etc.). Such differences in terms of structure and productivity in the families will be put forward within the framework of *LexRom*.

In terms of access to words, two primarily functionalities are foreseen (in a similar way to *Polymots*), namely, access by word-form and access by semantic cues.

First, by typing a **key word** ('olive') the user will obtain the equivalent families in all the existing languages. This may include families where a same stem produces several semantic clusters ('olive' and 'oil' in Spanish) or narrower families corresponding to a single semantic cluster ('olive' in French, but not 'oil'). Correspondences between equivalent words will be highlighted. Clusters of one family conveyed by other stems ('oil' in French) will be available by navigation through the lexical graph. The user will thus be able to browse from family to family and from a particular word to its equivalents.

Second, by entering **semantic cues**, the user will obtain a word in the same language as the one used to enter the conceptual items (e.g. 'tree', 'Mediterranian', 'robust' and 'peace' will display the 'olive' family because of the key word 'olive tree' obtained as a result of the query). A list of all the words in the same families in the other languages will also be displayed. Semantic cues, initially extracted from French, will be automatically translated for the other languages. They will provide a dynamic way to access to words.

Finally, other criteria will be proposed for searching, i.e. **productivity** of a family in a particular language or **specific affixes across languages**, to give two examples among others.

At the time of the writing this paper, *LexRom* is under development: 1 741 families for French, 190 for Spanish and 77 for Catalan have been gathered (about 25 000 words overall). Automatic acquisition with very large

[10] This example is a clear evidence that derivational morphology is very frequent in Romance languages, while other languages like English 'prefer' other morphological processes to create new words (compounding, composition, etc.).

[11] Lexical productivity stands for the number of derived forms of a stem. Interlingual contrastive examples may be explored with *LexRom,* e.g. the word *chaise* 'chair' has a single derived word in French *-chaisier-* while its Spanish equivalent *silla* generates up to twenty-three derived items.

corpora and other existing lexicons is needed to scale up the resource (which will be freely available under a Creative Commons licence).

5. JeuxDeMots

JeuxDeMots (Lafourcade, 2007) is a web-based associative game[12]where people are invited to play on various lexical and semantic relations between terms. A particular formatting of the lexical network with a associative dictionary tool is already available for users[13], allowing them to browse the network by jumping from term to term through relations (free association, hyperonym, hyponym, part-of, typical locations, typical subjects and objects for verbs, etc.). Semantic text analysis is the main application for exploiting this resource, however the use as a tool for providing help in the case of the 'tip of the tongue' phenomena is also fruitful.

5.1 A game on popular consensus

Semantic information is collected through non negotiated popular consensus. By consensus, we mean that when several players independently propose during a game the same association, they are memorized in the lexical network. The approach is non negotiated as players are playing without knowing until the result of the game with who they are playing, hence avoiding some interaction than would bias strongly the result. To ensure a system leading to quality and consistency of the database, it was decided that the validation of the relations given by a player should be made by comparison with those of other players. Practically, a relation is considered as correct if it is given by at least one pair of players, making v*alidation by pairs* a form of minimum filtering. This approach is similar to the one used by (von Ahn et al., 2004) for the indexation of images or more recently by (Lieberman et al., 2007) to collect common sense knowledge. As far as we know, this was never done in the field of the lexical networks. In NLP some other web-based systems exist, such as *Open Mind Word Expert* (Mihalcea and Chklovski, 2003) that aims to create large sense tagged corpora with the help of Web users, or *SemKey* (Marchetti et al., 2007) that exploits *WordNet* and *Wikipedia* in order to disambiguate lexical forms to refer to a concept, thus identifying a semantic keyword.

A typical game takes place between two players, in an asynchronous way, based on the concordance of their propositions. When a first player (A) starts a game, an instruction related to a relation type (synonyms, opposite, domains,...) is displayed, as well as a term T pseudo-randomly picked in a base of terms. Player A has then a limited amount of time to answer by giving propositions which, to his mind, correspond to the instruction applied to the term T. The number of propositions is limited inducing players not just type anything as fast as possible, but to choose amongst all answers he can think of. The same term, along the same instruction, is later proposed to another player B; the process being then identical. To increase the playful aspect, for any common answer in the propositions of both players, they receive a given number of points as a reward. The calculation of this number of points is crafted to induce both precision and recall in the feeding of the database. More precisely an answer that is very commonly given would not be much rewarding contrary to an original one. Thus, players have to deal with a double opposite constraints: trying to think like others (to have words in common) while being as original as possible (to get points).

At the time of the writing of this paper, more than 1 100 000 relations linking more than 230 000 terms have been collected. More than one million games (with a mean of 1 minute per game) have been played corresponding to approximately 17 000 hours (about 700 days) of cumulative play. The lexical resources produced with *JeuxDeMots* are freely available under a Creative Commons licence.

5.2 Tool and Evaluation

The question of evaluation the lexical network is difficult. Indeed, there is no comparable resources that could be used as a golden standard. Another way to tackle the problem is to devise a tool that could help people finding a word they have on the tip on the tongue. The success rate of the tool can serve as a rough evaluation of the underlying resource. AKI is such a tool, where people are invited to submit clues one after the other until the system proposes the expected terms, or fails. AKI is used as a tool for lexical access but in fact most people use it as a game where the goal is to challenge the guessing capabilities. Clues can take the form of words or a composition between a relation type and a term. For exemple, the clue *:isa* animal, states that the target word is an animal. Around 20 relations are available. Amongst other, AKI proposes: *:syn* for synonyms, *:antofor* antonyms, *:mat* for mater or substances (like *:mat* wood, the target term is made of wood), *:carac* for typical feature (like *:carac* white), *:part* for typical parts (like *:part* wheel), *:do* for typical action (like *:do* meow), etc.

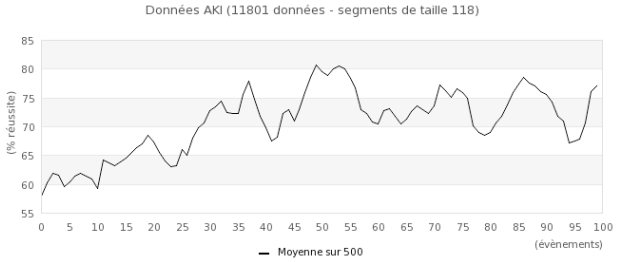
For example, a typical tip of the tongue game would be:

***:isa* boat**	**liner**
shipwrec	**Titanic**
torpedoing	**Lusitania**

Table 2: Typical AKI game (clues are on the left and answers made by AKI on the right)

[12] http://jeuxdemots.org
[13] http://www.lirmm.fr/jeuxdemots/diko.php

We evaluated the AKI performance for about 10 000 games. The overall performance is about 74% of success. We made an assessment of human performances on 200 games randomly taken from those played with AKI, and discover that under the very same conditions people have around 48% success.



6. Conclusion

LexRom and JeuxDeMots illustrate alternative ways from classic lexicographic work to create and enrich lexical resources. They both provide novel accesses and usages that are feasible only in the context of computer and user networks. Obviously, although the major part of the data is acquired automatically or by contributions, evaluation and validation of the resources with human contribution is essential to ensure the linguistic quality of the data. In this sense, NLP and lexicographic approaches converge undoubtedly. Still, the coverage of the resources, i.e. the amount of data obtained with automatic acquisition or collaborative contributions, remains a key issue of such novel approaches.

Last but not least, a number of open issues remain. As for *LexRom*, as the data is gathered mainly from corpora, the constitution of very large high quality corpora in different languages is crucial. At the time of writing this article we are working on such direction to be able to obtain more data. Likewise, theoretical aspects concerning lexical holes and polysemy deserve special attention in terms of language modeling. The *JeuxDeMots project* faces the same issues as previously mentioned concerning the coverage, but also concerning the qualitative evaluation. Some forthcoming research directions will include more common sens knowledge and the introduction of various non standard lexical relations. Collecting lexical information of very particular relations, either specialized or with too few possible answers, doesn't seem to be feasible with games. Thus, some contributive approaches with strong user incentives are still to be invented.

7. Acknowledgments

The authors would like to thank M. Zock as well as the reviewers of the paper for their valuable insights.

8. References

Atkins, B.T.S., Zampolli A. (1994). *Computational Approaches to the Lexicon*. Oxford: Oxford University Press.

Baayen, H.R., Piepenbrock, R. & Gulikers, L. (1995). *The CELEX Lexical Database* (CD-ROM). Linguistic Data Consortium, Philadelphia: University of Pennsylvania.

Briscoe, T., Carrol J. (1997). Automatic extraction of subcategorization from corpora. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*. Washington, DC, pp. 356-363.

Boitet, C., Mangeot, M. & Serasset, G. (2002). The PAPILLON project: cooperatively building a multilingual lexical data-base to derive open source dictionaries and lexicons. In N. Ide, G. Wilcock (eds.) *Proceedings of on Natural Language Processing and XML, COLING Workshop*. Taipei, Taiwan, pp. 9-15.

Calzolari, N., Choukri, K., Fellbaum, C., Hovy, E. & Fellbaum, C. (1999) *Multilingual resources. Chapter 1. Multilingual Information Management: current levels and future abilities*. Report commissioned by the US National Science Foundation and the European Commission's Language Engineering Office.

Clément, L., Sagot, B. & Lang, B. (2004). Morphology based automatic acquisition of large-coverage lexica. In *Proceedings of Fourth international conference on Language Resources and Evaluation (LREC-2004)*, Lisbon, Portugal, pp. 1841-1844.

Cristea, D., Forăscu, C., Răschip, M. & Zock, M. (2008). How to Evaluate and Raise the Quality in a Collaborative Lexicographic Approach. *In Proceedings of Sixth international conference on Language Resources and Evaluation (LREC-2008)*, Marrakech.

Fellbaum, C. (1998). *WordNet: an Electronic Lexical Database*. Cambridge, MA: MIT Press.

Fellbaum, C., Miller, G.A. (2003). Morphosemantic links in WordNet. In M. Zock, J. Carroll (eds.) *Les dictionnaires électroniques*. TAL vol. 44 (2), pp. 69-80.

Fort, K., Adda, G. & Cohen, K.B. (2011). Amazon mechanical turk: Gold mine or coal mine? *Computational Linguistics*, 37(2), pp. 413-420.

Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M. & Soria, C. (2006). Lexical Markup Framework (LMF). In *Proceedings of Fifth international conference on Language Resources and Evaluation (LREC-2006)*, Genoa, Italy.

Gala, N., Rey, V. & Zock, M. (2010). A tool for linking stems and conceptual fragments to enhance word access. In *Proceedings of the 7th international conference on Language Resources and Evaluation (LREC*), La Valetta, Malta.

Gala, N. (2011). Developing a lexicon of word families for closely-related languages. In *Proceedings of ESSLLI International Workshop on Lexical Resources* (WoLeR). Ljubljana, Slovenia.

Gasiglia, N. (2009). Evolutions informatiques en lexicographie: ce qui a changé et ce qui pourrait émerger. *Lexique 19*, pp. 224-298.

Grefenstette, G. (1998), The future of linguistics and lexicographers: will there be lexicographers in the year 3000? In T. Fontenelle, P. Hiligsmann, A. Michiels, A.

Moulin & S. Theissen (eds) *Proceedings of the Eighth EURALEX Congress*. Liège: University of Liège, pp. 25-41.

Heid, U. (2009). Aspects of lexical description for electronic dictionaries. Key note speech at *Electronic lexicography in the 21st century (ELEX-2009)*, Louvain, Belgium. http://www.uclouvain.be/en-271026.html.

Ide, N., Véronis, J. (1994). Machine Readable Dictionaries: what have we learned, where do we go? In *Proceedings of the post-COLING International Workshop on Directions of Lexical Research. Beijing, China.*

Jakubíček, M., Kilgarriff, A., McCarthy, D. & Rychlý, P. (2010). Fast syntactic searching in very large corpora for many languages. In *Proceedings of Workshop on Advanced Corpus Solutions, PACLIC 24.* Tohoku Univeristy, Japan.

Kilgarriff, A. (1997). Putting frequencies in a dictionary. *International Journal of Lexicography*, 10(2), 135-155.

Lafourcade, M. (2007). Making people play for Lexical Acquisition. In *Proceedings of the 7th Symposium on Natural Language Processing*. Pattaya, Thailand.

Lieberman, H., Smith, D.A. & Teeters, A. (2007). *Common Consensus: a web-based game for collecting commonsense goals, IUI'07*, Hawaii, USA.

Mihalcea, R., Chklovski, T. (2003). Open Mind Word Expert: Creating Large Annotated Data Collections with Web Users' Help. In *Proceedings of the EACL 2003 Workshop on Linguistically Annotated Corpora (LINC 2003)*, Budapest.

Marchetti, A., Tesconi, M., Ronzano, F., Rosella, M. & Minutoli, S. (2007). SemKey: A Semantic Collaborative Tagging System. In *Proceedings of WWW Conference*, Banff, Canada.

Navigli, R., Ponzetto, S. (2010). BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, 11-16 July 2010, pp. 216-225.

Navigli, R., Velardi, P. & Gangemi, A. (2003). Ontology learning and its application to automated terminology translation. *IEEE Intelligent systems*, 18(1), pp. 22-31.

Poesio, M., Kruschwitz, U. & Chamberlain, J. (2008). ANAWIKI: Creating Anaphorically Annotated Resources through Web Cooperation. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, 2008.

Véronis, J., Ide, N. (1990). Word Sense Disambiguation with Very Large Neural Networks Extracted from Machine Readable Dictionaries. In *Proceedings of 13th International Conference on Computational Linguistics (COLING'90)* Helsinki ,vol. 2, pp. 389-394.

Von Ahn, L.L.D. (2004). Labelling images with a computer game. In *Proceedings of ACM Conference on Human Factors in Computing Systems (CHI) 2004,* Vienna, Austria.

Vossen, P., (1998). *EuroWordNet: A Multi- lingual Database with Lexical Semantic Networks*. Dordrecht, The Netherlands: Kluwer.