Do we still Need Gold Standards for Evaluation?

Thierry Poibeau, Cédric Messiant

Do we still Need Gold Standards for Evaluation?

Thierry Poibeau and Cédric Messiant

Laboratoire d'Informatique de Paris-Nord
CNRS UMR 7030 and Université Paris 13
99, avenue Jean-Baptiste Clément
F-93430 Villetaneuse France
firstname.lastname@lipn.univ-paris13.fr

Abstract

The availability of a huge mass of textual data in electronic format has increased the need for fast and accurate techniques for textual data processing. Machine learning and statistical approaches have been increasingly used in NLP since the 1990s, mainly because they are quick, versatile and efficient. However, despite this evolution of the field, evaluation still rely (most of the time) on a comparison between the output of a probabilistic or statistical system on the one hand, and a non-statistic, most of the time hand-crafted, gold standard on the other hand. In order to be able to compare these two sets of data, which are inherently of a different nature, it is first necessary to modify the statistical data so that they fit with the hand-crafted reference. For example, a statistical parser, instead of producing a score of grammaticality, will have to produce a binary value for each sentence (grammatical *vs* ungrammatical) or a tree similar to the one stored in the treebank used as a reference. In this paper, we take the example of the acquisition of subcategorization frames from corpora as a practical example. Our study is motivated by the fact that, even if a gold standard is an invaluable resource for evaluation, a gold standard is always partial and does not really show how accurate and useful results are. We describe the task (SCF acquisition) and show how it is a typical NLP task. We then very briefly describe our SCF acquisition system before discussing different issues related to the evaluation using a gold standard. Lastly, we adopt the classical distinction between intrinsic and extrinsic evaluation and show why this framework is relevant for SCF acquisition. We show that, even if intrinsic evaluation correlates with extrinsic evaluation, these two evaluation frameworks give a complementary insight on the results. In the conclusion, we quickly discuss the case of other NLP tasks.

1. Introduction

The availability of a huge mass of textual data in electronic format has increased the need for fast and accurate techniques for textual data processing. Machine learning and statistical approaches have been increasingly used in NLP since the 1990s, mainly because they are quick, versatile and efficient.

However, despite this evolution of the field, evaluation still rely (most of the time) on a comparison between the output of a probabilistic or statistical system on the one hand, and a non-statistic, most of the time hand-crafted, gold standard on the other hand.

In order to be able to compare these two sets of data, which are inherently of a different nature, it is first necessary to modify the statistical data so that they fit with the hand-crafted reference. For example, a statistical parser, instead of producing a score of grammaticality, will have to produce a binary value for each sentence (grammatical *vs* ungrammatical) or a tree similar to the one stored in the treebank used as a reference (tree edit distances are rarely used).

There is thus a major bias in this classical evaluation scheme, which is nevertheless still the most widely used one in NLP. We take as an example the automatic acquisition of subcategorization frames (SCF) from corpora, since this task has been increasingly popular in the last few years and has produced a set of available and useful resources. We will not describe the basic techniques used for the automatic acquisition of SCF here, but we think that this example is relevant when discussing problems related to the gold standard approach for evaluation (see (Messiant and Poibeau, 2008) and (Messiant, 2008) for the description of our system; (Korhonen, 2002) or (Schulte im Walde, 2006) for other systems concerning different languages).

We will first describe the task (SCF acquisition) and show how it is a typical NLP task (section 2). We will then very briefly describe our SCF acquisition system (section 3) before discussing different issues related to the evaluation using a gold standard (section 4). Lastly, we adopt the classical distinction between intrinsic and extrinsic evaluation and show why this framework is relevant for SCF acquisition (section 5). We show that, even if intrinsic evaluation correlates with extrinsic evaluation, these two evaluation frameworks give a complementary insight on the results. In the conclusion (section 6), we briefly discuss the case of other NLP tasks.

2. SCF Acquisition as a Typical NLP Task

This paper takes the acquisition of lexical information from corpora as a typical task for NLP; the evaluation of the task (here the evaluation of data obtained from corpora) entails common problems shared by most NLP tasks.

It is well known that a dictionary, encoding accurate lexical knowledge, is a key component of most applications. Common electronic dictionaries can include structured data (e.g. hierarchies of semantic classes) with complex information (e.g. SCFs, selection restrictions). For example, the association of a list of SCFs with a given predicate is a key component of most syntactic parsers: these parsers need to have access (among other things) to the number and the nature of the arguments of the verb (NP, PP, infinitive clause, *etc.*) in order to be able to accurately analyze a sentence. However, a dictionary of predicative items (verbs, nouns and adjectives) including information about their SCFs is still

not available for most languages, including French. The automatic acquisition of such data from corpora, even if not perfect, largely reduces the time spent on the development of resources, especially when compared to a manual approach.

As for most linguistic questions, there is no well-established definition of what to include in a SCF, but everybody agrees that a SCF should minimally include the number and the type of the complements dependent from the verb (or from the predicative item considered, since adjectives and nouns can also govern a SCF). Most authors agree on the fact that complements should be divided between arguments and adjuncts but the distinction between these two categories is far from obvious. Some linguistic tests exist (can the complement be deleted without changing the meaning of the sentence? Can it be moved easily? Can it be pronominalized? etc.) but none of these tests is sufficient or discriminatory enough.

As outlined by Manning (Manning, 2003) "rather than maintaining a categorical argument / adjunct distinction and having to make in/out decisions about such cases, we might instead try to represent SCF information as a probability distribution over argument frames, with different verbal dependents expected to occur with a verb with a certain probability". For example, from the analysis of a large news corpora, one can observe that the French verb *venir (to come)* accepts the frame *PP[de (from)]* with a relative frequency of 59.1% whereas it accepts the frame *PP[à (to)]* with a relative frequency of 5%. This phenomenon is a kind of selectional "preference" of certain verbs for certain SCFs; the link with more semantic information remains to be done.

However, the evaluation of probability distributions is difficult, since it is by definition dependent from a given corpus. Hand-crafted dictionaries do not contain this kind of information. We need to investigate how existing dictionaries can be used and to what extent they can be considered as gold standard.

Of course, dictionaries are not the only possible gold standards for the evaluation of SCF acquisition: for example, large annotated corpora have also been used, especially for English (Korhonen, 2002). It is self-evident that a proper evaluation should take into account these various sources of information (dictionaries and annotated data). However, the comparison with a dictionary, considered as a gold standard, is the most frequently used evaluation framework (especially for languages other than English), so we will mostly focus on it.

This task is typical in that most NLP tasks are now based on stochastic or probabilistic approaches. We will briefly discuss in the last section a few other examples than the acquisition of SCF from corpora and we will show that the same questions arise then.

3. The SCF Acquisition System

The SCF acquisition system will be described very briefly here, as it only stands as an example to discuss the evaluation framework and the use of gold standards in evaluation. More detailed explanations can be found other publications (see (Messiant and Poibeau, 2008) (Messiant, 2008)).

The SCF acquisition system takes as input a large corpus and produces a list of frames for each verb that occurred enough in the corpus. Partial lists of SCF associated with verbs already exist for French (see next section) but our system is able to derive automatically information for a large number of verbs from a representative corpus, along with frequency information. One of our goal is to be able to quickly and automatically tune lexical information for a new corpus or a new domain, since it is well known that lexical information is largely dependent from the domain or the text genre. The system also gathers statistical information that is useful for stochastic parsers or subsequents processes (like the inference of lexical classes from the SCF distributions).

Below is a typical lexical example obtained from the corpus analysis.

```
:NUM: 05204
:SUBCAT: s'abattre : SP[sur+SN]
:VERB: S'ABATTRE+s'abattre
:SCF: SP[sur+SN]
:COUNT: 420
:RELFREQ: 0.882
:EXAMPLE: 25458;25459;25460;25461;25462
```

The entry is related to the French verb *s'abattre* (to crash down), that takes a prepositional complement introduced by the French preposition *sur*. Other information corresponds to the relative frequency of this frame for the verb (0.882) and links to several examples.

4. Evaluating our Results against a Gold Standard

Even if statistical approaches are now widely used in NLP, most of the evaluations done so far for lexical acquisition are based on a comparison against a hand-crafted gold standard. The first experiments have been done on languages for which such a gold standard was available (English (Korhonen, 2002), German (Schulte im Walde, 2006)). In such a case, it is possible to check if a given verb has received a list of "correct" SCFs (i.e. the acquired SCF is also registered in the gold standard), if some are missing (i.e. a SCF is present in the gold standard, but not in the acquired data) or over-generated (i.e. a SCF present in the acquired data, not in the gold standard).

In this section, we first describe existing resources for French, we examine how reliable they are and discuss their use as a Gold Standard.

4.1. Existing resources for French

Even if there is currently no comprehensive dictionary for French (i.e. a dictionary containing an exhaustive list of SCFs for each verb), a number of resources can be used as a basis for the evaluation of our system. The most relevant ones are quickly described below.

- Dicovalence (`http://bach.arts.kuleuven.be/dicovalence`) is a dictionary for verbs (Van Den Eynde and Mertens, 2006). It includes manually defined SCF frames for 3,700 simple French verbs

(8,000 entries, no idiom). Those entries are modelled using the Pronominal Approach (Van Den Eynde and Blanche-Benveniste, 1978): for each syntactic slot, Dicovalence specifies the paradigm of associated pronouns, which describes intentionally all possible lexicalizations.

- Lexicon-Grammar (`http://infolingu.univ-mlv.fr/`) is a hand-crafted dictionary developed by a team of researchers led by Maurice Gross (Gross, 1994). The Lexicon-Grammar (LG) for French includes syntactic information for a large number of French words (including verbs, nouns and adjectives – the resource includes 5,000 entries for simple verbs along with a large number of verb compounds) encoded through binaries features; Information in LG is dispatched through these features and must be translated into a format which is more amenable for use by NLP systems (Claire et al., 2005). Only a part of the resource is publicly available.

- Lefff (`http://alpage.inria.fr/catalogue.fr.html#Lefff`) (Sagot et al., 2006) is a syntactic lexicon that distinguishes two levels of lexical description: the intensional level, which includes syntactic information and the extensional level, which is the (highly redundant) list of inflected form generated automatically from the intensional level. Lefff contains over 114,000 intensional entries; it has been compiled from various sources and has not been fully validated.

- TreeLex (`http://erssab.u-bordeaux3.fr/article.php3?id_article=150`) is a subcategorization lexicon of verbs which has been automatically extracted from the Paris 7 treebank (Kupść, 2007). It contains about 2,000 verbs with their subcategorization frames and information about the frame frequencies.

- TLFI (*Trésor de la Langue Française Informatisé*) is the most complete resource available for French. This electronic dictionary has been derived from a classical paper dictionary that includes etymological, morphological, syntactic and semantic information for most French words. Since this resource is not a machine-readable dictionary it cannot be used directly but has to be manually translated, in order to infer formalized SCF from classical entries. Examples are then especially important but the translation process is not completely obvious since implicit information has to be made explicit sometimes.

4.2. How gold is the gold standard?

All these dictionaries are good starting points for evaluation, but none can be used directly.

The first thing that should be noted concerns the coverage of the resource to be used. TLFI is the most comprehensive resource for French. It has been fully validated and is publicly available with examples. On the other hand, TreeLex has been derived from a one million-word corpus and thus has a low coverage. However, TLFI is not directly usable and must be translated in order to be used for evaluation (as noted above). It has been developed using a corpus of French classical literature and is sometimes not completely adapted to modern French.

We should notice that TLFI is not the only one that needs to be translated in order to be usable. Even a resource like LG, which is an electronic resource intended to be used in computational systems, has to be translated in order to obtain explicit SCFs (Claire et al., 2005). A resource like Dicovalence is encoded using the Pronominal Approach (Van Den Eynde and Blanche-Benveniste, 1978), which makes it not so easy to use: sets of pronouns have to be translated into possible surface realization; some key elements are sometimes missing (for example, prepositions included in PP are not always specified in Dicovalence, especially for location phrases), etc.

Finally, some dictionaries are not fully available (especially LG). Some others have not been fully validated (Lefff) or have been automatically extracted from medium size annotated corpora, which means that their coverage is rather small (TreeLex). None of them has productivity information, except TreeLex (and in this last case, productivity has been computed from a corpus on 1,000,000 word only, which is small to get relevant productivity information). Note that the projection of an existing resource on a specific corpus to get productivity information is far from obvious, since lots of ambiguities have to be evaluated (several SCFs can be applied for a given sentence most of the time).

All these aspects should be taken into consideration when designing the evaluation. However, it is often difficult to get a gold standard that exactly fits with the task and the format of the data obtained automatically. One must keep these points in mind for the evaluation.

4.3. The need for a more thorough evaluation

Despite all these caveats, we performed a classical evaluation, by comparing our results with a gold standard. Since TLFI si the most comprehensive dictionary available, we chose to perform the evaluation by comparing our results against TLFI used as a gold standard.

We randomly chose a set of twenty verbs that were heterogeneous enough in terms of semantic and syntactic features. The system extracted twenty SCFs from the corpus for these verbs with a mean of 4.47 frames per verb.

We then calculate type precision (i.e. evaluation against SCF types found in some dictionary or some of the input data), token recall (i.e. manually annotate some of the input data for SCFs and see how many occurrences the system analyses correctly) and F-measure (harmonic mean of precision and recall). It yields 0.79 precision, 0.55 recall and 0.65 F-measure. These results are similar to those obtained by (Korhonen et al., 2000) despite the apparent differences between French and English and the absence of a predefined list of frames for French. The only comparable previous experiment for French, (Chesley and Salmon-Alt, 2006), obtained slightly different results (0.86 precision and 0.54 recall) but their overall F-measure (0.66) is quite similar to ours. We assume that the slight variation in precision and recall can be explained by the nature of the corpus and the choices in the filtering method.

However, we found it rather difficult to "evaluate the evaluation". It is not clear whether a 0.65 F-measure is enough to be practically usable. Of course, evaluating against a gold standard is not intended to say anything about the practical usefulness of a resource but this information would be more interesting for most people.

We have also observed that this kind of evaluation suffers from several biases. The gold standard includes several SCFs that are not found in our results. Some of these frames correspond to old or very specific constructions that occurred in classical French but are no more present in modern French. On the other hand, acceptable frames found in the corpus were not present in the gold standard (domain-dependent or new syntactic constructions). All of this leads to a lower precision and recall compared to a manual or more practical evaluation.

Moreover, the acquired resource is not fully comparable to a hand-crafted resource: gold standards are generally based on a strong distinction between arguments and adjuncts, whereas statistical approaches give a weight corresponding to the strength of the link between the verb and the complement. Most of the time, parsed trees obtained by syntactic parsers are plausible and can be perfectly fine for most NLP tasks, even if not always completely comparable to manually annotated corpora (Bod, 2007).

Hand-crafted data used as a gold standard (e.g. the *TLFI*) do not contain any information about productivity of the different SCFs. Since this element is a key point for stochastic parsers, they obtained a worse performance when no statistical information is provided with the corresponding SCFs. However, frequency information cannot be provided with a hand-crafted gold standard and therefore, cannot be evaluated.

We then tried to use different resources as a gold standard. This is not always possible without a lot of work, especially to translate the different resources in a common format. We then obtained various scores for precision, recall and F-measure but since these scores depend from the gold standard, it does not give very usable results. Since dictionaries (used as gold standards) do not have the same coverage, are not developed from the same basis (classical literature, news, technical documents) and are not encoded in the same way, this evaluation did not lead to any comparable results.

However, we do not claim here that evaluating against a gold standard is completely ineffective. It is largely admitted that a gold standard with a manual verification still remains an invaluable resource for evaluation, especially for recall (to check what has been missed by the acquisition process).

However, in order to practically evaluate a resource, two main approaches are interesting and complementary: 1) ensuring that the gold standard is as comprehensive as possible (intrinsic evaluation) and 2) evaluating through a practical task (extrinsic evaluation).

5. Intrinsic *vs* Extrinsic Evaluation

In this section, we describe two different ways of evaluating practical results. Our proposal is not new, since it corresponds to the classical distinction made by (Jones and Gallier, 1996) between intrinsic evaluation (evaluation of the resource – or of the task – for itself) and extrinsic evaluation (evaluating the resource by integrating it in a practical application).

5.1. Intrinsic evaluation: Making the Gold Standard as Comprehensive as Possible

As shown above, no resource provides a really satisfactory gold standard for French. Our first experiments have been based on TLFI (and other existing dictionaries), but a more comprehensive resource should be built by comparing the different resource existing for French, merging their respective SCF and cross-validating the results against a representative corpus.

This approach is the one described in (Korhonen, 2002). In this experiment, two large dictionaries for English (ANLT and COMLEX) are merged and a large corpus is annotated. The evaluation is made against this set of cross-validated resources, thus offering more accurate results. The same should be made for French, but merging resources is a time-consuming task, especially in our case, since existing dictionaries are based on very different theoretical backgrounds. A few research groups have elaborated a multi-year, multi-institutions project to achieve this goal, but the result will not be ready before several years.

Moreover, (Briscoe, 2001) notes that (semi-)manually developed lexicons tend to show high precision but disappointing recall (even when merging several large, already existing dictionaries to get a reference). It is often difficult to detect what is missing in a given dictionary and it is largely ineffective to manually check these dictionaries. Moreover, it is well known that no resource can ever be complete, since lexical information depends on genre, domain and discourse types (Biber, 1988). Automatic acquisition paired with lexical tuning thus remains the most promising approach to overcome these shortcomings (Wilks et al., 1996).

It is then relevant to add a more practical approach to the classical evaluation framework, by taking into account the performance of the acquired lexicon in a practical task. In the next section, we show how SCF acquisition can be evaluated through an information extraction task.

5.2. Extrinsic Evaluation: Does it Correlate with Intrinsic Evaluation?

The usefulness of extrinsic evaluation has been demonstrated by several authors (among others (Dorr et al., 2005), from which this title is inspired). The question is then: does this other kind of evaluation correlates with intrinsic evaluation?

We have shown that the comparison with a gold standard is not always the best way to evaluate a given tool. From this point of view, the evaluation of SCF acquisition is not an isolated case: Rens Bod, while evaluating a parser, claims that "it is well known that any evaluation on hand-annotated corpora unreasonably favours supervised parsers. There is thus a quest for designing an evaluation scheme that is independent of annotations" (Bod, 2007). He proposes to evaluate against a practical task (machine translation in this case).

An alternative way of evaluating a lexical resource is thus to integrate it in a practical application. For example, a set of verbs with SCFs acquired from a representative corpus has been integrated in a parser by Carroll and Briscoe (Briscoe and Carroll, 1997). Then, they evaluate the contribution of SCFs for parsing. They obtain better results when the SCFs are integrated into their parser, compared to when the parser is purely non-lexicalized.
Practical tasks such as Information Extraction also provide interesting ways of measuring the quality of a resource. Information Extraction largely depends on SCF information: in order to extract structured information, it is necessary to know what elements are dependent from a given predicate and what is their role in the action expressed by the predicate.
We have shown in several experiments (e.g. (Poibeau et al., 2002), (Poibeau, 2007)) that automatic acquisition from corpora allows one to find specialized items that are not mentioned in a general domain resource. These elements make up from 30 to 45% of the useful information. Therefore, acquisition from corpora increases recall.

- Some SCFs are not registered in existing lexical databases since they are domain-specific. These are most of the time crucial for the task (*<N0> faire une OPA sur <N1>*, *<N0> initiate a takeover on <N1>* for example) and can be successfully acquired from the corpus. For most of the Information Extraction frameworks that we have developed, even when these elements were not numerous, their productivity was high. Acquiring them from the corpus increases recall significantly.

- Some complements are considered as adjuncts and are not registered in existing lexical databases for French. However, most of these adjuncts are relevant and are captured as such by statistical approaches, which mainly rely on corpus productivity. For example, in the case of acquisition verbs (*buy*, *get*, *acquire*, ...), the information about price is nearly always included in our corpus (*<N0> buy <N1> for <Amount>*), whereas this information is not mentioned, or mentioned as an adjunct, in standard dictionaries.

Since this information is not included in the gold standard, it is clear that the use of lexical information acquired from a corpus gives more accurate results than the use of a handcrafted dictionary. This conclusion correlates with the results obtained from the gold standard but gives a practical proof of the interest of the task.
Note however that most of the domain-specific elements are idioms or multi-word expressions (as opposed to simple verbs). Acquisition should thus also focus on these elements, which are very poorly described in most handcrafted dictionaries. Up to now, most SCF acquisition modules are concerned with simple verbs and ignore multi-word expressions, which are more difficult to grasp.
On a more theoretical ground, this approach gives a new basis for the distinction between arguments and adjuncts: it shows that probability distribution is a relevant factor for the issue. It also reflects the relations between words, idioms and constructions (Croft and Cruse, 2004) and, therefore, the fact that it is hard to evaluate them separately. Practical tasks require to deal with all these levels at the same time, whereas they are artificially split up when performing an intrinsic evaluation.

6. Conclusion: What about other tasks?

In this paper, we have taken the example of the acquisition of subcategorization frames from corpora as a practical example. Our study was motivated by the fact that, even if a gold standard is an invaluable resource for evaluation, a gold standard is always partial and does not really show how accurate and useful results are.
A gold standard generally provides an interesting basis for the comparison of systems against the same set of data, or for the comparison of the evolution of the performance of the different versions of a system performing a certain task. This is especially true when one work on a new language and do her first experiments without any idea of SCF types, with no annotated data and no tasks ready where to plug the acquired SCFs. In this case, using a dictionary as a gold standard seems to be a good starting point, but only a starting point.
Our observations are not new, since they correspond to the classical distinction between intrinsic and extrinsic evaluation. However, extrinsic evaluation is rarely done for SCF acquisition since it is labour-intensive and require to have a practical application available. Moreover, one has to elaborate a clear evaluation protocol, in order to make a difference between errors due to the lexical component and errors due to the application itself.
However, several authors have shown that extrinsic evaluation yields interesting results for a large number of tasks, either data-oriented (e.g. lexical acquisition (Poibeau et al., 2002)), module-oriented (e.g. parsing, (Bod, 2007)) or user-oriented (e.g. automatic summarization, information extraction, machine translation, (Dorr et al., 2005)).

7. Acknowledgement

This research is part of the ANR MDCO project CroTal. Cédric Messiant's PhD is partially funded through a DGA Grant.

8. References

Douglas Biber. 1988. *Dimensions of Register Variation : A Cross-linguistic Comparison*. Cambridge University Press, Cambridge.

Rens Bod. 2007. Is the End of Supervised Parsing in Sight? In *Association for Computational Linguistics*, pages 400–408, Prague.

Ted Briscoe and John Carroll. 1997. Automatic Extraction of Subcategorization from Corpora. In *5th ACL Conference on Applied Natural Language Processing*, pages 356–363, Washington, DC.

Ted Briscoe. 2001. From Dictionary to Corpus to Self-Organizing Dictionary: Learning Valency Associations in the Face of Variation and Change. In *Corpus Linguistics Conference*, Lancaster University.

Paula Chesley and Susanne Salmon-Alt. 2006. Automatic Extraction of Subcategorization Frames for French. In *Language Resources and Evaluation Conference (LREC)*, Genoa (Italy).

Gardent Claire, Bruno Guillaume, Guy Perrier, and Ingrid Falk. 2005. Maurice Gross' Grammar Lexicon and Natural Language Processing. In *2nd Language and Technology Conference*, Poznan.

William Croft and Alan Cruse. 2004. *Cognitive Linguistics*. Cambridge University Press, Cambridge.

Bonnie Dorr, Christof Monz, Stacy President, Richard Schwartz, and David Zajic. 2005. A Methodology for Extrinsic Evaluation of Text Summarization: Does ROUGE Correlate? In *ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 1–8, Ann Arbor, Michigan. Association for Computational Linguistics.

Maurice Gross. 1994. Constructing Lexicon-Grammars. In *Computational Approaches to the Lexicon*, pages 213–263, Oxford. Oxford University Press.

Karen Sparck Jones and Jean Gallier. 1996. *Evaluating Natural Language Processing Systems: An Analysis and Review*. Oxford University Press, Oxford.

Anna Korhonen, G. Gorrell, and D. McCarthy. 2000. Statistical filtering and subcategorization frame acquisition. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, Hong Kong.

Anna Korhonen. 2002. *Subcategorization Acquisition*. Ph.D. thesis, University of Cambridge.

Anna Kupść. 2007. Extraction automatique de cadres de sous-catégorisation verbale pour le français à partir d'un corpus arboré. In *Conférence Traitement Automatique du Langage Naturel*, Toulouse.

Christopher D. Manning. 2003. Probabilistic syntax. In R. Bod, J. Hay, and S. Jannedy, editors, *Probabilistic Linguistics*, pages 289–341. MIT Press.

Cédric Messiant and Thierry Poibeau. 2008. LexSchem: A Large Subcategorization Lexicon for French Verbs. In *Language Resource and Evaluation Conference (LREC)*, Marrakech.

Cédric Messiant. 2008. A Subcategorization Frames Acquisition System for French Verbs. In *Association for Computational Linguistics (ACL, Student Research Workshop)*, Columbus, Ohio.

Thierry Poibeau, Dominique Dutoit, and Sophie Bizouard. 2002. Evaluating Resource Acquisition Tools for Information Extraction. In *Language Resources and Evaluation Conference (LREC)*, Las Palmas.

Thierry Poibeau. 2007. Semantic annotation: Mapping Text to Ontologies. *International Journal of Metadata, Semantics and Ontologies*, 2(2).

Benoît Sagot, Lionel Clément, Eric de La Clergerie, and Pierre Boullier. 2006. The Lefff 2 Syntactic Lexicon for French: Architecture, Acquisition, Use. In *Language Resource and Evaluation Conference (LREC)*, Genoa.

Sabine Schulte im Walde. 2006. Experiments on the Automatic Induction of German Semantic Verb Classes. *Computational Linguistics*, 32(2):159–194.

Karel Van Den Eynde and Claire Blanche-Benveniste. 1978. Syntaxe et Mécanismes Descriptifs : Présentation de l'approche pronominale. *Cahiers de Lexicologie*, 32:3–27.

Karel Van Den Eynde and Piet Mertens. 2006. *Le dictionnaire de valence Dicovalence : manuel d'utilisation*. Manuscript, Leuven.

Yorick Wilks, Brian Slator, and Louise Guthrie. 1996. *Electric Words: Dictionaries, Computers, and Meanings*. The MIT Press, Cambridge, MA.