Maurice Gross' grammar lexicon and Natural Language Processing

Claire Gardent[♢], Bruno Guillaume[♠], Guy Perrier[♡], Ingrid Falk[♣]

[♢]CNRS/LORIA
[♠]INRIA/LORIA
[♡]University Nancy 2/LORIA
[♣]CNRS/ATILF
Nancy, France
FirstName.LastName@loria.fr

Abstract

Maurice Gross' grammar lexicon contains an extremly rich and exhaustive information about the morphosyntactic and semantic properties of French syntactic functors (verbs, adjectives, nouns). Yet its use within natural language processing systems is still restricted. In this paper, we first argue that the information contained in the grammar lexicon is potentially useful for Natural Language Processing (NLP). We then sketch a way to translate this information into a format which is arguably more amenable for use by NLP systems.

1. Maurice Gross's grammar lexicon

Much work in syntax concentrates on identifying and formalising general syntactic rules that are thought to be valid of a large class of words. Typically, Chomsky's transformation rules describe systematic relations between syntactic structures. And more recently, the lexical rules of e.g., Lexical Functional Grammar systematically describes a pair of syntactic categories deemed to hold of a given class of words.

But as Chomsky himself observed (Chomsky, 1965), these generalisations are subject to strong lexical constraints. Given a specific word, the question whether or not a given generalisation applies needs to be answered. Or in other words, a full description of the syntax of a language implies not only the identification of general syntactic rules but also, and equally importantly, a detailed specification of which word requires, accepts or forbids the application of which syntactic rule. This is what Maurice Gross' work on the grammar lexicon (Gross, 1975) sets out to achieve for the French language.

Maurice Gross' grammar lexicon is a systematic description of the syntactic properties of the syntactic functors of French namely, verbs, predicative nouns and adverbs.

This lexicon is organised in groups of tables, each group containing the syntactic descriptions associated with a given syntactic category (verb, support verb construction, nouns, etc.).

Further, in a group, a table denotes a specific syntactic construction (sometimes two) and groups together all the lexical items entering in that construction. For instance, the first table in the group of tables for verbs groups together all the verbs which can take besides a subject, an infinitival complement but not a finite or a nominal one.

Finally, for each item in a given table, a set of columns further specify the syntactic properties of that item either by adding information about its arguments or by identifying a number of transformations the basic subcategorisation frame associated with the table can undergo.

At present, the grammar lexicon is most developed for verbs and verbal locutions. For so called "simple verbs", 5 000 verbs have been described over a total of 15 000 verb usages (Gross, 1975; Boons et al., 1976a; Boons et al., 1976b). Further, 25 000 verbal locutions are also described as well as 20 000 locutions using "être" (to be) or "avoir" (to have) (Gross, 1989).

2. The need for electronic lexicons in Natural Language Processing

For natural language systems, knowledge acquisition is a main bottleneck. We concentrate here on the morphosyntactic knowledge associated with verbs and show that the information contained in the grammar lexicon is highly relevant for NLP systems. Specifically, we argue that the grammar lexicon contains (at least) two types of information that is of use for NLP namely, subcategorisation and alternation information.

Subcategorisation. The grammar lexicon contains detailed and exhaustive information about subcategorisation that is, about the number and the type of arguments a verb can take. Specifically, the information that can be recovered from the LADL tables includes for each verb usage described:

- one or more basic subcategorisation frame(s) consisting of a list of arguments
- and detailed morpho-syntactic information about both verb and arguments including among others:
  - for the verb : information about the verb type (defective,normal,u-verb), about the auxil-

We would like to thank Eric Laporte and the Institut d'électronique et d'informatique Gaspard-Monge for making some of the LADL tables available to us in eletronic format. We would also like to thank the Contrat Plan Etat Région : Ingénierie des Langues, du Document et de l'Information Scientifique, Technique et Culturelle for partially funding the research presented in this paper.

iary used to construct composed tenses (être or avoir), about tense concordancy constraints on verbal arguments, etc.

- for nominal arguments : information about animacy, number, selectional restrictions, pronominalisation, restriction on the determiner, etc.
- for prepositional arguments : information about the type (e.g., locative) and about the value of the preposition used
- for sentential arguments : information about the mood (declarative, infinitive, subjunctive), the control structure of the verb (subject vs object control), possible verb instantiations, etc.

As is shown by current and recent research work in NLP, this detailed subcategorisation information is an essential component in enhancing the linguistic coverage and the accuracy of NLP systems. Indeed because many current computational theories of syntax project syntactic structures from the lexicon, parsers based on these theories must have access to accurate and comprehensive information concerning the number and the types of arguments taken by syntactic functors and in particular, by verbs.

More specifically, (Briscoe and Carroll, 1993) shows that half of parse failures on unseen data test results from inaccurate subcategorisation information in the ANLT dictionary while (Carroll and Fang, 2004) demonstrates that for a given domain, using an HPSG (Head Driven Phrase Structure Grammar) enriched with detailed subcategorisation information improves the parse success rate by 15%.

Since in many applications, parsing often occurs early in a pipeline of several NLP modules, accurate information about the subcategorisation properties of syntactic functors is a key component in ensuring quality output for these applications. As demonstrated by (Han et al., 2000) for instance, it is a key factor in achieving good quality machine translation.

Detailed subcategorisation information is also essential in ensuring a good basis for semantic construction and thus for semantic processing in general. Consider the following example from (Carroll and Fang, 2004) for instance:

(1) I'm thinking of buying this software too but the trial version doesn't seem to have the option to set priorities on channels

To correctly compute the basic functor/argument structure of this sentence, it is essential that the underlined prepositional phrases be recognised not as modifiers but as arguments of the corresponding verbs. Although such a basic functor/argument structure is a far cry from reconstructing the meaning of a sentence, it is a basic ingredient in constructing it. Thus for instance, (Jijkoun et al., 2004) shows that extracting syntactic relations between entities in a text, rather than using surface-based patterns, substantially increases the number of factoid questions answered by a question answering system.

Alternations. Another type of information contained in the LADL tables which is highly relevant for NLP systems is the information about verb alternations it contains[1] that is, about the possible deletions and movement the arguments of a syntactic functor can undergo. For instance, a verb can be specified as (dis)allowing the following alternations :

- passive *Le chat mange la souris/La souris est mangée par le chat*
- reciprocal *Luc flirte avec Léa/Luc et Léa flirtent*
- locative alternation *Les fautes pullulent dans ce texte/Ce texte pullule de fautes*
- source alternation *Un paradoxe résulte de cette situation/De cette situation résulte un paradoxe*
- inchoative form *Jean sonne la cloche/La cloche sonne*
- support verb construction *Jean crie/Jean pousse un cri*
- body part possessor ascension alternation *Jean imite l'attitude de Marie/Jean imite Marie dans son attitude*

For the English language, Beth Levin has carried out an extensive study of such alternations whose aim was to identify semantic verb classes (Levin, 1993). The driving intuition is that syntactic variations reflect semantic ones. The methodology used by Beth Levin is then to identify for each verb the set of alternations this verb participates in and to define verb classes on the basis of this alternation information : verbs that (dis)allow the same set of alternations are grouped into a common class.

Because it provides a sound empirical and theoretical basis for verb classification, Levin's work has had a major impact in computational linguistics. It is used in particular as a basis for VerbNet (Kipper et al., 2000), an electronic verb lexicon with syntactic and semantic information for roughly 2 500 English verbs. The essential point is that Levin's classes (or rather the intersective Levin's classes defined in (Dang et al., 1998)) provide the appropriate level of abstraction for describing the syntactic and semantic properties of verbs. As a result, it becomes possible to develop highly factorised verb lexicons thus avoiding maintenance and consistency problems. As (Kipper et al., 2000) show, the resulting resource provides a detailed description both of the syntactic alternations associated with a given verb and of its basic lexical semantics namely its thematic grid and a reasonably abstract decompositional semantics. And in the same way that accurate detection of syntactic dependencies can improve question/answering system, a resource that contained detailed and exhaustive information about thematic grids and lexical semantics is an important ingredient in supporting accurate semantic processing.

3. Existing electronic lexicons

Although it is now clear that extensive and detailed computational lexicons are needed to improve the cover-

[1] It is usual in the literature to distinguish between alternations and redistributions, the former being less generally applicable than the former. For simplicity and because the border between the two phenomena remains fuzzy, we englobe here both of them under the term "alternation".

age and the accuracy of NLP systems, few such lexicons actually exist.

For the English language, COMLEX Syntax (Macleod et al., 1994) contains detailed subcategorisation information for 38 000 words of which 6 000 are verbs and VerbNet describes 4 000 verbs senses using 191 semantic verb classes and 52 subcategorisation frames.

For the French language on the other hand, a number of electronic lexica are available but these are not related to subcategorisation and verb semantics. Thus, the LEFFF lexicon (Lexique des Formes Fléchies du Franais is extensive and contains 5 000 verbs with 200 000 forms but it only contains flectional information (Clément et al., 2004). Similarly, the morphological word lists extractable from the ATILF databases (Dendien and Pierrel, 2003), MulText (Ide and Veronis, 1994) and ABU (Association des Bibliophiles Universels ABU,) are restricted to morphosyntactic information. Regarding alternations, (Saint-Dizier, 1999) describes the alternations of French verbs but is limited to 1 000 verb forms.

In sum, there is neither an extensive and available electronic lexicon for French which describes basic subcategorisation frames nor one which describes the alternations of verbs.

4. The grammar lexicon as a basis for a computational verb lexicon

As we saw in section 2., Gross' grammar lexicon contains detailed and exhaustive information about both subcategorisation and alternations. Moreover, the grammar lexicon has been digitised by the Laboratoire d'Automatique Documentaire et Linguistique (LADL) and is now partially available under an LGPL licence. Hence the grammar lexicon information is available for use in digitised format and can be used as a basis to create a lexical resource appropriate for use by NLP systems. To achieve this goal however, several changes are required in the way the information is structured and formatted. More specifically:

- The information pertaining to a given verb must be collected and put together into one or more lexical entries.
- The data structures and the linguistic categories used to represent the information must be compatible with usual practice in computational linguistics.
- The format of the data must be compatible with state of the art practice in data formatting.

In what follows, we concentrate on the first point namely the grouping of all information pertaining to a given verb within one single lexical entry. The next two points are briefly addressed at the end of the section where we pinpoint the important issues arising in that area and suggest directions for future work.

In NLP applications, linguistic information is standardly retrieved from an electronic lexicon where each word is associated with its linguistic properties. In the COMLEX Syntax dictionary for instance, each entry describes the syntactic properties of a given adjective, noun or verb usage. Furthermore, these properties are organised as a nested set of feature value structures.

More generally, a standard, reasonably theory and application neutral way to represent lexical information consists in (i) associating with each word one or more lexical entries and (ii) describing the content of these entries using recursive feature structures that is, sets of attribute-value pairs where values can be (negated or disjoint) atoms, strings or feature structures.

Thus one first step in turning the existing grammar lexicon in a "meta lexicon" usable by various NLP modules and applications consists in converting the content of the tables into a set of lexical entries, each entry associating with a given word usage, the set of linguistic properties assigned to it by the grammar lexicon. We report here on some preliminary work done in that direction and illustrate the process by showing how table 1 can be converted into a set of lexical entries.

The general idea is to process each table one after the other and to create for each verb occurring in each table a set of lexical entries as described by the content of the grammar lexicon. For a given table, the general conversion procedure can be described as follows.

1. For each verb V mentioned in table T, create a lexical entry associating V with the basic subcategorisation frame associated with T.
2. Enrich each lexical entry created in step 1, using the content of table T columns for V.

A subcategorisation frame is defined by a list of atoms (e.g., A_0 V A_1) representing the verb and its arguments) and by a list of atoms/feature structure pairs specifying the feature values associated with each of these atoms. So for instance, the basic subcategorisation frame associated with Table 1 is noted as indicated below where the U feature pertains to Harris U verb class, CAT denotes the part of speeech, MODE the verb mood and CONTROLEUR indicates the controller of the infinitival complement in this case the subject.

```
a0 v a1
v:=[u=+]
a1:=[cat=p,mode=inf,controleur=a0]
```

The processing for each verb of the table columns may then either enrich this specific entry or create new ones (for the same verb). So for instance, the processing of Table 1 for the verb *traîner* enrich its basic subcategorisation frame as follows:

```
{a0 => {hum => -, nc => +},
  v => {particule_post => "l\'a", cat => u,
        concTemps => -, passivable => +,
        prep => [\'a], aux => [\^etre]},
 a1 => {vc => [pouvoir,savoir,devoir],
        tc => [pass\e,pr\esent,future],
        cliticisable => +, cat => p,
        mode => [inf,ind,subj],
        controleur => a0, optional => 1}}
```

Furthermore, certain columns of the table indicate that a given transformation is applicable to the basic subcategorisation frame of *traîner* so that further lexical entries are created e.g.:

```
{a0 => {hum => -, nc => +},
  v => {particule_post => "la",
        cat => u, concTemps => -,
        passivable => +,
        prep => [a], aux => [etre]},
  a1 => {cat => sp, hum => +}}
```

Due to space restrictions, we cannot detail here the content of the procedures yielding the above lexical entries. In essence, the processing proceeds in two steps. First, the effect of the columns is manually identified (creation or enrichment of a lexical entry) and translated into an and-or graph representing the various subcategorisation frames described by the table. Second, an algorithm is defined which (i) creates the subcategorisation frames described by the and-or graph and (ii) instantiates for each verb in the table the various subcategorisation frames associated by the table to that verb. The approach markedly differs from that described in (Hathout and Namer, 1998) in that the creation of the and-or graph involves a detailed "manual" reinterpretation of the table headings and of their interdependencies.

To be widely usable, a resource must conform to general linguistic and computational usage. Linguistically, feature names and categories should be used which "make sense" to the widest possible audience. To this end, we intend to make use of the catalogues proposed by MulText, EAGLES and more recently by the Lexical Markup Framework ISO (TC37/SC4) standard. The latter in particular, provides a high level model for representing data in lexical resources and thus guarantees a maximum of interoperability with multilingual computer applications. Computationally, it is important to use a language which supports efficient and generalised processing. XML is in this respect a natural candidate as it is a *de facto* standard supporting information structuring, structure checking and querying.

5. Conclusion

The work reported here is preliminary. Current and future work concentrates on extending the approach to the full set of tables currently available. This involves (i) abstracting away from the table descriptions the general principles underlying the structuring of the LADL tables, (ii) agreeing on a set of features and feature values to be used and (iii) developing the algorithms necessary to convert the content of the grammar lexicon into an NLP friendly meta lexicon usable by different people for different applications.

6. References

Association des Bibliophiles Universels ABU. Dictionnaire des mots communs. Conservatoire National des Arts et Metiers.

Boons, J.-P., A. Guillet, and C. Leclère, 1976a. *La structure des phrases simples en français. I : Constructions intransitives*. Droz, Gen ève.

Boons, J.-P., A. Guillet, and C. Leclère, 1976b. La structure des phrases simples en français. ii : Classes de constructions transitives. Technical report, Univ. Paris 7.

Briscoe, E. and J. Carroll, 1993. Generalised probabilistic lr parsing for unification-based grammars. *Computational Linguistics*.

Carroll, J. and A. Fang, 2004. The automatic acquisition of verb subcategorisations and their impact on the performance of an hpsg parser. In *Proceedings of the 1st International Joint Conference on Natural Language Processing (IJCNLP)*. Sanya City, China.

Chomsky, N., 1965. *Aspects of the theory of syntax*. The MIT Press.

Clément, L., B. Sagot, and B. Lang, 2004. Morphology based automatic acquisition of large-coverage lexica. In *Proceedings of LREC'04*. Lisbonne.

Dang, H.T., K. Kipper, M. Palmer, and J. Rosenzweig, 1998. Investigating regular sense extensions based on inters ective levin classes. In *Proceedings of COLING-ACL98*. Montreal, Canada.

Dendien, J. and J.-M. Pierrel, 2003. Le trésor de la langue française informatisé : un exemple d'informatisation d'un dictionnire de langue de référence. *Traitement Automatique des Langues*, 44(2).

Gross, G., 1989. *Les constructions converses du français*. Droz, Gen ève.

Gross, M., 1975. *Méthodes en syntaxe*. Hermann.

Han, C., B. Lavoie, M. Palmer, O. Rambow, R. Kittredge, T. Korelsky, and N. Kim, 2000. Handling structural divergences and recovering dropped arguments in a korean/english machine translation system. In *Proceedings of the Association for Machine Translation in the Americas*. Berlin/New York: Springer Verlag.

Hathout, N. and F. Namer, 1998. Automatic construction and validation of french large lexical resources: Reuse of verb theoretical linguistic descriptions. In *First International Conference on Language Resources and Evaluation, Granada, Spain*.

Ide, N. and J. Veronis, 1994. Multext: Multilingual text tools and corpora. In *Proceedings of COLING 94*. Kyoto.

Jijkoun, V., J. Mur, and M. de Rijke, 2004. Information extraction for question answering: Improving recall through syntactic patterns. In *COLING-2004*.

Kipper, K., H. Trang Dang, and M. Palmer, 2000. Class based construction of a verb lexicon. In *Proceedings of AAAI-2000 Seventeenth National Conference on Artificial Intelligence*. Austin TX.

Levin, B., 1993. *English verb classes and alternations: a preliminary investigation*. Chicago University Press.

Macleod, C., R. Grishman, and A. Meyers, 1994. Comlex syntax: Building a computational lexicon. In *Proceedings of COLING '94*.

Saint-Dizier, P., 1999. Alternation and verb semantic classes for french: Analysis and class formation. In *Predicative forms in natural language and in lexical knowledge bases*. Kluwer Academic Publishers.