From the corpus to the lexicon: the example of data models for verb subcategorization.

Paula Chesley, Susanne Salmon-Alt

From the corpus to the lexicon: the example of data models for verb subcategorization

PAULA CHESLEY
Linguistics Department, University at Buffalo, USA & ATILF-CNRS, France
pchesley@buffalo.edu

SUSANNE SALMON-ALT
ATILF-CNRS, France
susanne.alt@loria.fr

Abstract

This paper describes the integration of corpus-based syntactic subcategorization frames and correlated semantic information into a large-scale, cross-theoretically informed lexical database for French (Romary et al. (2004)). This database is the first to implement the Lexical Markup Framework (LMF), an international initiative towards ISO standards for lexical databases (ISO TC 37/SC 4). The subcategorization frames have been acquired via a dependency-based parser (Bick (2003)), whose verb lexicon is currently incomplete with respect to subcategorization frames. Therefore, we have implemented probabilistic filtering as a post-parsing treatment using the binomial distribution. Building on our discussion of what semantic information, e.g., participant roles, to include in the database, we describe how we plan to exploit our findings on subcategorization frames to derive this information via unsupervised learning techniques.

1 Introduction

This paper describes the integration of corpus-based syntactic subcategorization frames and correlated semantic information into a large-scale, cross-theoretically informed lexical database for French (Romary et al. (2004)). The Morphalou database is freely downloadable (http://www.atilf.fr/morphalou) and is the first to implement the Lexical Markup Framework (LMF), an international initiative towards ISO standards for lexical databases (ISO TC 37/SC 4). We thus discuss how the recommended LMF data structures for syntactic and semantic information guide our acquisition of subcategorization frames of verbs from annotated corpora, as well as the representation of the syntax-semantics interface in wide-coverage lexical databases. Because automatic subcategorization extraction for French is a nascent field, we have chosen to initially concentrate on extracting subcategorization information of verbs. Subsequently, we aim to extend this research to other parts of speech that also exhibit subcategorization phenomena.

At present there is no lexical database for French that encodes lexical syntactic and semantic information to the extent that the current research plans to do so. For example, the French component of EuroWordNet does not contain any subcategorization frame or argument structure information. Furthermore, establishing a standard lexical database format ensures that information in this format will remain functional and exploitable for many years to come. Thus, each component of the current undertaking is an essential step for Natural Language Processing (NLP) for French.

As Briscoe and Carroll (1997) note in a study on English, incorrect syntactic subcategorization information is responsible for approximately half of parsing errors like incorrect prepositional attachments. Assuming similar figures for French, any French parser would benefit from subcategorization information that the Morphalou database will contain. In addition, monolingual and bilingual dictionaries note subcategorization information. The manual alternative to automatic extraction of subcategorization frames for dictionaries is of course expensive, time-consuming, and potentially difficult to re-use.

Additionally, Gildea and Jurafsky (2002) note that domain-specific semantic information is employed in spoken dialogue and information extraction systems, but that there is yet a lack for general semantic information such as that of Fillmore's (1976) semantic frames. Having such related information, the authors note, would allow verbs of a common frame, such as *send* and *receive* from the TRANSFER frame, to share the same semantic roles, thus aiding in a question-answering system in which one verb is in the question, while another is in the response. Furthermore, Nasr (2004) notes that, inter alia, lack of argument structure information in an annotated learning corpus for French could constitute a reason for which his dependency parser did not fare as well for French as it did on the English test corpus containing such information.

Following existing initiatives in the modeling of syntactic and semantic lexical knowledge (Genelex, ISLE/MILE, etc.), the integration of subcategorization frames into the LMF data model occurs at several levels which are both syntactic and semantic in nature. On the syntactic side, one central data structure characterizes a lexical entry, that of a set of syntactic constructions. This construction set corresponds to a set of frames observable in a corpus for a given entry with a given sense. Each syntactic construction is further described by a set of syntactic positions, i.e., the syntactic category (nominal phrase, subordinate clause, etc.) and syntactic function (subject, direct object, etc.) of each element subcategorized for by the verb. Further information about the LMF data model is found in section 2.

Since semantic information is not directly observable in corpora for French, we aim to use corpus-observed subcategorization information to infer semantic knowledge to be incorporated in the lexical database. Our experiment is based on a corpus we have created from Frantext, an online literary French database. The subcategorization frames have been acquired via a dependency-based parser (Bick (2003)), whose verb lexicon is currently incomplete with respect to subcategorization frames. Therefore, we have implemented probabilistic filtering as a post-parsing treatment using the binomial distribution. This sort of dual treatment constitutes a technique shown to be successful in subcategorization frame filtering for English (Brent (1992), Brent (1993), Manning (1993), Briscoe and Carroll (1997)). Our study differs from that of Bourigault and Frérot (2005), who are undertaking research in subcategorization of prepositional phrases (PPs). That is, in addition to PPs, the present work seeks to extract subordinate clauses, the impersonal *il* subject, and direct and indirect objects in complete frames. Section 3 details the subcategorization frame extraction process.

Determining a verb's semantic traits to be integrated into the Morphalou database relies on work in linking in theoretical linguistics. Once decided upon, we advance the hypothesis that semantic information can be uncovered via unsupervised learning of observable linking phenomena and surface cues subcategorization frames in corpora. However, even from correct subcategorization frames, the proper number of semantic arguments for the verb is difficult to obtain, since some participant roles are optional (Koenig et al. (2003)). We can, nevertheless, infer some information about participant roles, since for at least some verbs and some subcategorization frames there is a direct correlation to a semantic argument. For example, the preposition *vers*, 'toward' in a PP complement, maps to a semantic *locative* or *goal* argument. However, not all surface cues will yield dependable semantic information. The semantic aspects of the database are discussed in section 4.

2 LMF, an international standard for lexical databases

Lexical structures can classically be considered according to the way they organize the relation between words and senses. On the semasiological view, senses are considered as subdivisions of the lexical entry, whereas on the onomasiological view, words are considered as ways of expressing concepts. Of these views, the former allows an exhaustive survey of lexical content for a given language.

In particular, it corresponds to the basis for any classical editorial, or print, dictionary, and also underlies, at least implicitly, most existing NLP lexicons. From a theoretical perspective, the internal structure of a lexical entry can be configured through different layers. In a two-layered approach, the `/form/` and `/sense/` layers are anchored to the Saussurian definition of a linguistic sign and are related to the basic notions of *signifier*, the sound pattern of a lexical entry, and *signified*, the corresponding concept. The syntactic behavior of the lexical unit is thus systematically subjacent to its semantic description. This notion is currently being implemented in the LMF and is being developed in the ISO TC 37/SC 4 as a future standard for the representation of lexical resources (Francopoulo et al. (2004)). Accordingly, the LMF core model is organized as a hierarchical structure built upon the following components:

- the `/lexicalDatabase/` component, which gathers all information related to a given lexicon;
- the `/globalInformation/` component collecting metadata such as version number, contributors, updates made, etc.;
- a `/lexicalEntry/` component, which corresponds to the elementary lexical unit in a lexical database;
- a `/form/` component providing access to surface properties, i.e., phonological and graphical realizations as well as grammatical properties such as inflectional features;
- one or more `/sense/` components, which currently organize the lexical entry. These components can be repeated, in the case of homonymy, and further divided into sub-senses in the case of polysemy.

Furthermore, following general principles of the linguistic annotation scheme design stated in Ide and Romary (2003), the LMF provides a mechanism for combining the components of the basic data model with elementary descriptors, or data categories. Data categories reflect basic morphosyntactic concepts (e.g. `/partOfSpeech/`, `/grammaticalNumber/`, `/grammaticalCase/`, etc.). They are stored and managed independently from the hierarchical structure of the data model. Proceeding in this way allows for recording language-specific properties independently of structural properties of the linguistic layers to be described. For instance, the data category `/grammaticalGender/` holds two values for French, `/masculine/` and `/feminine/`, and three values for German, `/masculine/`, `/feminine/` and `/neuter/`. In order to share data categories within the community, the ISO/TC 37 deploys an online data-category registry[1] for use in conjunction with the other standardization activities. The future LMF standard as such does not aim to provide a specific list of data categories to be used for lexical descriptions. Doing so would be far too complex, given the potential variety of applications. It is thus expected that implementers will systematically refer to the ISO/TC 37 data category registry to find the proper descriptive background for their individual needs.

Finally, the LMF provides mechanisms to translate the combination of the core model and data categories into an isomorphic XML pivot structure. The implementers might then chose to express their own combination of a core model and data categories in an LMF-XML "dialect". For example, it is possible to implement a given data category such as `/grammaticalGender/` as an XML element rather than an attribute, or by renaming it as `/gen/`, `/gender/` or `/genre/`. Crucially, such a proprietary XML dialect must be able to be mapped unambiguously to the LMF-compatible XML pivot structure in order to ensure proper standardization.

[1] This registry is accessible at `http://syntax.inist.fr`.

2.1 Extending the LMF to syntax and semantics

In its current state, the syntactic extension of LMF essentially covers syntactic realizations of argument structures for entries with predicative senses, especially verbs. The researchers involved have not yet come to a consensus on the ensemble of components to be used. In the present work, our description thus proceeds from the concrete LMF structures, i.e., a model for data structures directly observable in a corpus, to the more abstract. The most concrete data structures in the syntactic component are at the level of the syntactic dependent, the syntactic realization of semantic argument, the data category for which is `/syntacticArgument/`. A syntactic dependent is minimally described by the following data categories:

- `/syntacticFunction/`, having basic values such as `/subject/`, `/directObject/` and `/prepositionalObject/` which might be able to be refined with user defined data categories for language-specific phenomena;
- `/syntacticConstituent/`, describing the syntactic category of the argument, e.g. `/nounPhrase/`, `/prepositionalPhrase/`, `/subordinateClause/`, etc.;
- `/syntacticIntroducer/`, allowing the user to record, in case of prepositional phrases or subordinate clauses, the preposition or the complementizer. This data category can of course be extended to languages that make use of other introducers like postpositions for Korean.

In addition, each syntactic dependent has a /`semanticRestriction`/ data category. This data category can contain participant roles as values as well as other lexical semantic information. More data categories for the syntactic dependent level may be added according to further research.

The LMF allows for a recursive description of subordinate clauses in terms of a set of syntactic dependents, thus providing a simple way of encoding various morphosyntactic constraints on subordinates such as mood, tense and co-indexation of subject or object. It is also possible to add examples or occurrence frequencies at various levels of granularity. This can prove useful for illustrating a particular syntactic dependent with a corpus example or to count the occurrences of a particular realization of a syntactic dependent.

Building upward, the first level of abstraction is the syntactic construction, with the data category `/syntacticConstruction/`, which represents a subcategorization frame. We define a subcategorization frame as a set of syntactic dependents, realized simultaneously by a predicative lexical entry. The introduction of this component corresponds to the need of an anchoring point for lexical information about the whole construction, e.g., the auxiliary verb for past participle forms, or constraints on ordering and/or the simultaneous realisation of different syntactic arguments. More fundamentally, syntactic constructions are the basic data structures on which syntactic alternations and transformations are effectuated. Depending on the degree of extensionality of the lexicon, the user might decide to explicitly encode the entire range of surface syntactic constructions (including, for example, passive constructions), or to encode only canonical constructions (`/canonicalConstruction/`) to be associated with grammatical rules if the lexicon is to be used in conjection with a parser.

A further abstraction is the grouping of those abstract constructions into classes sharing the same syntactic behavior with respect to alternations and transformations. In the LMF, this component is referred to as a verb's lexical class and represents the crucial point of the syntax-semantics interface in the model. It can be understood as a class of verbs similar to Levin's (1993) seminal work for English, or a table number from Maurice Gross' (1975) tables denoting the syntax-semantics interface for French verbs. This and other aspects of the LMF architecture are given in the example of a lexical entry in appendix B.

Having established the outline of the LMF in its current state, we turn now to the practical concerns of incorporating lexical information for French into the Morphalou database.

3 Subcategorization frame extraction

Extensive work in subcategorization frame extraction for French has been carried out by Bourigault and Frérot (2005). These experiments have been effectuated on large-scale corpora of approximately 200 million words. The focal point of these experiments is subcategorization of PPs; thus, subordinate clauses, impersonal subjects, and nominal and adjectival attributes are not discussed in detail. In addition, PPs are treated independently of other syntactic dependents – in fact, all syntactic dependents are treated independently of each other. In effect, this research does not obligatorily correlate a subcategorized PP to an attested frame: if the PP is dependent on another dependent that is absent, the PP might be misanalyzed as a dependent for which the verb subcategorizes rather than a modifier. An example of this phenomenon is given in (1).

(1) a. Jean a reçu un message de (la part de) la sécrétaire.
"Jean received a message from the secretary."

b. Jean a reçu de la confiture (pour son anniversaire).
"Jean received jam (for his birthday)."

In (1a), *de* is a preposition introducing a subcategorized PP, while in (1b) it is an indefinite article in the direct object. If we do not take into account other syntactic dependents – in this case, the direct object of (1a), we risk misparsing the direct object of (1b) as a PP. However, considering syntactic dependents independently of each other might prove sufficient most of the time. In addition, examining co-occurrences of syntactic dependents could induce errors due to subcategorization frame information that is too fine-grained. We simply felt it more judicious to err initially on the side of caution than to be potentially obliged to change our frame extraction methodology.

The present work on subcategorization frame extraction is based on a corpus of 115 verbs that we created from the online literary database Frantext. Our corpus comes from various genres between the years 1850 and 2000, such as treaties and novels, and excludes theatre and poetry, since these genres can yield statistically higher percentages of non-canonical subcategorization realizations than prose. With the query tool that Frantext provides, we have also excluded most occurrences of the causative construction, since it can change the argument structure, and thus the subcategorization frames, of a verb. Given these restrictions, we randomly chose 200 occurrences each of 115 verbs in the TSNLP for French to be parsed. A list of these verbs is given in appendix A. We used the VISL parser (Bick (2003)), a dependency-based parser whose lexicon is partially complete with respect to subcategorization frames[2]. Although in the parser analysis some dependent positions are more reliable than others we have chosen to weight all dependent positions equally and to let the filtering choose the correct subcategorization frames, since presumably filtering will also correctly label those which the parser does. We thus use the parser almost as a chunker that divides a sentence into phrases.

After the parsing stage, we effectuate a probabilistic filtering treatment that makes use of the binomial distribution. Dual treatments using this filtering method have proven extremely successful for subcategorization frame extraction for English (Brent (1993), Manning (1993), Briscoe and Carroll (1997)); for example, precision rates for Brent (1993) vary from 96% to 100% according to the frame. Let a *cue* be an initial frame we receive from the parser, without knowing whether it is indeed a frame.

[2]For a demonstration of the state of the art of the parser the reader may consult `http://visl.hum.sdu.dk/visl/en/parsing/automatic/trees.php`.

The binomial distribution in this application examines the difference between the number of times a particular cue occurs with a given verb and the number of total times the latter appears in the corpus. The greater this difference, the less likely it is that the cue is an actual frame. Let m be the total number of occurrences of a verb in the corpus, n be the number of co-occurrences of the verb with the cue, and B_f the estimated upper bound that the verb that does not subcategorize for the frame f appears nevertheless with f.We make the null hypothesis that the verb does not subcategorize for the cue. The upper bound on the probability that the hypothesis is false given all cues is the following (Manning (1993)):

$$\sum_{i=n}^{m} \frac{m!}{i!(m-i)!} {B_f}^i (1 - B_f)^{m-i}$$

Typical confidence levels are empirically set between .02 and .05, below which the cue is considered an actual frame. In the present work we have set the confidence level at .02.

Note that the binomial distribution supposes a known rate B_f, which is in effect the error rate for each cue. Since Manning (1993) examines 19 frames, he establishes this rate empirically for each of them. Brent (1994) details a method to establish the rate B_f automatically which we have adopted in the current work. In brief, this method consists in examining every occurrence of a frame with every verb above a certain number of occurrences, which we have currently fixed at 50. From these occurrences we construct a histogram based on the number of co-occurrences of cues and the verbs with a sufficient amount of corpus attestations. We look for a binomial distribution toward the lower end of the histogram that signals the false cues f. The average in this distribution is a proper estimation of the rate of false cues B_f. We refer the interested reader to Brent (1994) for an in-depth discussion of the method of finding B_f.

3.1 Evaluation

Once definitive results of this endeavor are established, we can seek to augment the Morphalou database with semantic information that is not only in accordance with theoretical linguistics but also easily exploitable for other NLP applications. We plan to evaluate subcategorization frames of verb types manually against a gold standard and the completed portion of the parser lexicon. Currently we have initial results and are examining what resources to use as a gold standard for a large-scale evaluation of our subcategorization frame extraction. Here are certain results, which we note in contrast to the initial parser results:

- **diriger**, 'to direct'. The parser does not include the frame `Sub V DO PP[`$_{vers}$ `'toward']`, as a frame for this verb, although the Collins Robert English-French dictionary includes it as such.
- **donner**, 'to give'. Our work indicates that `Sub V DO PP[`$_{\grave{a}}$ `'to']` is a frame for this verb. This is the standard ditransitive frame and the parser lexicon includes this frame.
- **courir**, 'to run'. Our experiment supposes the subcategorization frame `Il`$_{imp}$ `V PP[`$_{\grave{a}}$`]`. This frame does not exist in the parser lexicon, although the *Trésor de la langue française informatisé* indicates that it is indeed a subcategorization frame for this verb.
- **arriver**, 'to arrive'. Our experiment supposes the subcategorization frame `Il`$_{imp}$ `V PP[`$_{de}$ `'from']`. In fact, this frame does not exist. A frame with the same surface elements does however exist: `Il`$_{i}mp$ `V CMP[`$_{de}$ `(to)]` ("Il arrive de pleuvoir en été", "It can happen that it rains in summer"). This error could be due to an error in the parser. However, in our informal survey of the results, it appears that frames with the impersonal subject appear more often as part

of subcategorization elements than they should be. We might have to lessen our confidence level for this construction.

Currently bilingual dictionaries appear to have the most explicit subcategorization information. The *Trésor de la langue française informatisé* also appears to have a good amount of subcategorization information, and thus both of these resources could serve as gold standards in our evaluation process.

4 Inferring semantic information from subcategorization frames

Gildea and Jurafsky (2002) use the hand-labeled FrameNet database to build a classifier to discern 18 semantic roles, many of which are included in theoretical research on participant roles. This supervised learning technique is currently unavailable to us, as no hand-annotated corpus of semantic information currently exists for French. There is however large-coverage hand-annotated semantic information available in Gross (1975), in the form of multiple tables, albeit somewhat limited in nature. The traits *locative*, *proposition*, and *human* are consistently given in the tables. Gardent et al. (2005) are currently undertaking research to determine whether the tables are feasibly exploitable in their current format. However, subcategorization extraction must first be ensured in order to guarantee the proper linking bewteen semantic arguments and their syntactic realizations. After discussing what semantic information should be included in the Morphalou database, we address another possibility for obtaining this information, that of combining work in theoretical linguistics on linking and unsupervised learning techniques based on distributions of surface phenomena in the syntactic dependents of verbs in our corpus.

4.1 What semantic information and why?

The current lack of consensus as to what semantic information should be included in the Morphalou database reflects a similar dilemma in theoretical work in lexical semantics and linking. The issue of how the information encoded in a lexical item maps to a surface realization is no trivial issue and cannot be discussed here in great depth. However, as Koenig et al. (2003) note, most scholars are in agreement that "the syntactic structure of many sentences is mostly or entirely determined by the information about situation participants in lexical entries of verbs" (p. 69) (cf. Koenig and Davis (2001)). We thus contend that there is an essential step between a lexical entry and its syntactic realization that contains lexical semantic information; i.e., information about participant roles, that ought to be incorporated into a lexical database.

One way in which to achieve this mapping between semantic arguments onto syntactic dependents is to introduce a set of participant roles, such as *agent*, *patient*, etc. However, Koenig and Davis (2001) note that a principal drawback of this approach is that it cannot itself determine the number and types of participants required. These decisions must be left to linguists, whose opinions could well vary on the matter. Despite this shortcoming, we feel this approach, or variants of it based on current work on argument structure and linking, merit to be examined as a possible way in which to encode semantic information in the Morphalou database.

As much as possible, we would like the Morphalou database to respect a balance between the following concerns in regards to lexical semantic phenomena:

1. Intuitive conceptualization for non-linguists;
2. Linguistic accuracy;
3. A theory-neutral linguistic account.

It is worth noting that points 1 and 2 can at times appear as conflicting goals. For example, the semantic trait *human* is most likely more intuitive to non-linguists than the participant roles of *agent*, *experiencer*, *patient*, or *beneficiary* which a human can fulfill. However, the semantic trait *human* cannot be considered a participant role, while *locative* and *proposition* can be thought of as such. Participant role information should be abstract in nature; i.e., a term used to describe semantic information should not denote an entity existing in the world but rather a linguistic concept. Participant roles seem a more accurate description in light of productive phenomena such as metaphor and metonymy[3]. Additionally, participant roles are implemented as lexical semantic information in FrameNet.

4.2 Automatically acquiring semantic information

As opposed to the manually annotated resources with semantic information described in the tables of Gross (1975), the method we outline for obtaining participant roles has the advantage of being directly exploitable as soon as the evaluation of our subcategorization work is carried out. Clearly manually developed resources constitute a useful gold standard that should be exploited, but recall and precision rates of automatic extraction of subcategorization frames from them have yet to be established. Additionally, the tables do not employ all participant roles, nor do they make the distinction between linguistic concepts and concrete denotations in their semantic information; recall from the previous section that one semantic trait they employ is *human*.

A consensus concerning the ideal number of participant roles to distinguish has not yet been reached in theoretical linguistics. Therefore, we offer the following list as a first proposition of realistic participant roles to be automatically extracted given our corpus of subcategorization frames:

- agent;
- patient;
- location[4];
- instrument;
- beneficiary;
- experiencer;
- proposition.

Certain participant roles will be easier to extract than others. We thus sketch the surface cues and linking distributions which will aid us in extracting the participant roles before discussing possible methods for extracting this information.

In French the *agent* role rarely occurs as a direct or indirect object, barring of course the causative construction. Additionally, functional linguistics indicates that the *agent* role exhibits a strong preference for a syntactic realization of subject (Gildea and Jurafsky (2002)), although this is not always the case. What's more, agents and experiencers tend to be humans, and the distribution of humans in syntactic dependent realizations can be uncovered in using named entities. These facts and other distributional

[3]Markert and Nissim (2003) note that of 1,000 country names examined manually in the BNC, between 171 and 186 of country names constitute metonymical readings (17.1 - 18.6%), as opposed to 737 literal readings (73.7%). This percentage of metonomical uses is clearly significant for country names. A study confounding all named entities might yield similar results. However, it is not sure whether this percentage would still remain significant if all syntactic realizations of argument structure elements are examined.

[4]It remains to be seen if the *location* participant role can be further subdivided into the roles *goal*, *source*, and *destination*.

properties of agents, experiencers, and patients in French must be examined in further detail in order to extract these particpant roles on the basis of syntactic realizations of a verb's semantic arguments.

For the participant roles *beneficiary*, *location* and *proposition*, surface cues such as named entities, prepositions, and clitic realizations aid more in distinguishing these participant roles than for the *agent*, *experiencer*, and *patient* roles. As mentioned in section 1, certain prepositions such as *vers*, 'toward', only map to one participant role, that of *location*[5], when they represent heads of subcategorized PPs. Prepositions like *à* and *de*, respectively 'to' and 'from' or 'of', are far too common and do not map to a particular participant role, but many other prepositions, such as *chez*, 'at the house or establishment of' and *derrière*, 'behind', also demonstrate a direct mapping to a participant role. It is worth noting that these prepositions can also take a metaphorical, non-spatial sense. However, if the participant role changes due to this metaphorical usage, it does not seem probable that verbs subcategorize for these metaphorical senses of these prepositions, while they can and do for the concrete, spatial senses. Similarly, the *instrument* participant role in French is perhaps most often seen with a subcategorization realization of a PP headed by the preposition *avec*, 'with', and propositions are often introduced by the complementizers *de* and *que*, 'that'.

The object indirect in French is most often realized with the preposition *à*. As previously mentioned, this preposition cannot be directly mapped to a particular participant role. However, the indirect object can be cliticized into the unambiguous indirect object clitics *lui* and *leur* (some indirect object clitics also have the same form as direct object clitics). Of the seven participant roles noted above, the indirect object most often informs us about the *beneficiary* participant role.

After surface cues and linking distributions have been established, we can begin the bootstrapping process. Since this research is yet in its elementary stages, we present a simple sketch of how this bootstrapping could take place. Surface cues seem a promising direction. For example, we can say for every verb that subcategorizes for a preposition that maps only to a locative participant role, that the verb takes a locative participant role. The same might well true for a verb that subcategorizes for the preposition *avec*, 'with', and the instrument participant role. In addition, co-occurrence rates of unambiguous indirect object clitics and the verb can be examined, perhaps even in using the binomial distribution with a different confidence level than what we use in the current work, to see whether the verb takes the *beneficiary* partipant role. If we assume that the percentage of realizations of agent, experiencer, and patient roles realized as having the semantic trait *human* vary, a sample set of verbs with known argument structures can be examined for co-occurrence data of realizations with this trait via named-entity recognition. The verbs with similar co-occurrence data could be attributed the same participant roles as the verb with which it shares the most similar co-occurrence rate of *human* realizations.

5 Conclusions and perspectives

This work demonstrates the current state of the LMF structure and the content of the Morphalou database. Crucially, the theoretical work on the structure of the LMF is independent from our work on the content of the Morphalou database for French and will be able to be used for any language. We discuss how theoretical work in linguistics and lexicology influences our choices of structure and data categories. We also show the flexibility of the formats in regards to cross-linguistic diversity and mappings of other formats to the LMF. For example, users of the LMF can choose to use a subset of the data categories proposed in section 2.

The content of the Morphalou database is both syntactic and semantic, and we have detailed not only

[5]This participant role might be confounded with the more abstract role of *goal*.

how we are automatically extracting subcategorization frames for our test corpus of 115 verbs, but also how we plan to derive semantic information, i.e., participant roles of verbs, in the database. This content is similar to the semantic and syntactic information available in the English FrameNet, and we have illustrated our reasoning for selecting this semantic information in section 4.1.

After the evaluation of our subcategorization frame extraction, we principally plan to exploit these frames to derive semantic participant roles to be incorporated in the database. In so doing, we must examine the possibility of more surface cues that can aid in the detection of participant roles, as well as the possibility that current surface cues proposed may be erroneous. In the previous section, we propose that the unambiguous indirect object clitics *lui* and *leur* aid in determining beneficiary roles. However, we can think of at least one example, a psychological verb with the impersonal *il* subject, in which these clitics represent an experiencer rather than a beneficiary:

(2) Il lui plaît de voir sa cousine.
"It pleases him-IO to see his cousin."

Constructions that are contrary to our hypotheses must be examined for inherent patterns as well as frequency rates and productivity. We could also institute a potential default linking system for each frame, e.g., subjects could map to the partipipant role *agent*, according to frequency data of participant roles in a small sample corpus.

References

[Bick2003] E. Bick. 2003. A CG & PSG Hybrid Approach to Automatic Corpus Annotation. pages 1–12. Corpus Linguistics, Lancaster.

[Bourigault and Frérot2005] D. Bourigault and C. Frérot. 2005. Acquisition et évaluation sur corpus de propriétés de sous-catégorisation syntaxique. Actes des 12èmes journées sur le Traitement Automatique des Langues Naturelles.

[Brent1992] M. Brent. 1992. Robust Acquisition of Subcategorization Frames from Unrestricted Text: Unsupervised Learning with Syntactic Knowledge. Master's thesis, Johns Hopkins University, Baltimore, MD.

[Brent1993] M. Brent. 1993. From grammar to lexicon: Unsupervised learning of lexical syntax. *Computational Linguistics*, 19:243–262.

[Brent1994] M. Brent, 1994. *Surface Cues and Robust Inference as a Basis for the early Acquisition of Subcategorization Frames*, pages 433–470. MIT Press, Cambridge.

[Briscoe and Carroll1997] T. Briscoe and J. Carroll. 1997. Automatic Extraction of a Subcategorization from Corpora. pages 356–363. Proceedings of the 5th ACL Conference on Applied Natural Language Processing.

[Fillmore1976] C. Fillmore. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, 280:20–32.

[Francopoulo et al.2004] G. Francopoulo, M. George, and M. Pet. 2004. Data categories in lexical markup framework or how to lighten a model. LREC Workshop "A registry of linguistic data categories within an integrated language resources repository area".

[Gardent et al.2005] C Gardent, B. Guillaume, G. Perrier, and I. Falk. 2005. Maurice gross' grammar lexicon and natural language processing. Proceedings of the 2nd Language and Technology Conference.

[Gildea and Jurafsky2002] D. Gildea and D. Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28:3:245–288.

[Gross1975] M. Gross. 1975. *Méthodes en syntaxe: régime des constructions complétives*. Hermann, Paris.

[Ide and Romary2003] N. Ide and L. Romary, 2003. *Encoding Syntactic Annotation*, pages 281–96. Kluwer, Dordrecht.

[Koenig and Davis2001] J.-P. Koenig and A. Davis. 2001. Sublexical modality and the structure of lexical semantic representations. *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, 24:71–124.

[Koenig et al.2003] J.-P. Koenig, G. Mauner, and B. Bienvenue. 2003. Arguments for Adjuncts. *Cognition*, 89:67–103.

[Levin1993] B. Levin. 1993. *English Verb Classes and Alternations*. Chicago UP.

[Manning1993] C. Manning. 1993. Automatic Acquisition of a Large Subcategorization Dictionary from Corpora. pages 235–242. Proceedings of the 31st ACL.

[Markert and Nissim2003] K. Markert and M. Nissim. 2003. Corpus-based metonymy analysis. *Metaphor and Symbol*, 18:3:245–288.

[Nasr2004] A. Nasr. 2004. Analyse syntaxique probabiliste pour grammaires de dépendances extraites automatiquement. Habilitation à diriger des recherches, Université Paris 7.

[Romary et al.2004] L. Romary, G. Francopoulo, and S. Salmon-Alt. 2004. Standards going concrete: from LMF to Morphalou. COLING workshop.

A The 115 verbs in the corpus

aborder	croire	lire	regretter
accepter	croître	livrer	représenter
acheter	decider	maintenir	requérir
agir	démarrer	manger	réserver
aider	devenir	marcher	restaurer
aimer	devoir	marier	rester
aller	dire	mentir	rêver
apercevoir	diriger	mettre	savoir
apparaître	diviser	montrer	séparer
appeler	donner	offrir	signer
apprendre	dormir	ouvrer	sommer
arriver	durer	ouvrir	sortir
asseoir	écrire	paraître	sucrer
avertir	entendre	parler	suer
avoir	entreprendre	participer	suffire
avouer	entrer	partir	suivre
boire	espérer	passer	supposer
causer	étayer	penser	taire
cesser	être	permettre	terminer
combattre	exceller	persuader	tomber
commencer	faillir	plaire	toucher
comparer	faire	pleuvoir	transférer
comprendre	falloir	prendre	travailler
connaître	fontionner	présenter	trouver
constituer	hésiter	prononcer	venir
contrer	indiquer	proposer	vivre
convaincre	intéresser	provoquer	voir
courir	interroger	raconter	vouloir
craindre	laisser	recevoir	

B An example of a lexical entry in the Morphalou database

This format is the current implementation of the LMF. Note that the semantic information in the /`semanticRestriction`/ data category is not in conjunction with what is noted in section 4. This discrepancy is due to the fact that as of yet the /`semanticRestriction`/ information is that contained in the manually annotated tables of Gross (1975). In addition, researchers are still deciding upon what information to use in regards to the /`lexicalClass`/ data category. We limit ourselves to one example subcategorization frame of an entry due to space constraints.

```
<lexicalEntry id="" lemma="alarmer">
    <grammaticalCategory>verb</grammaticalCategory>
       <sense id="1" glose="to trouble" example="Que max parte ennuie Ida"
        source="LADL_table_4">
            <constructionSet source="LADL_table_4">
                <syntacticConstruction exampleConstruction="Max alarme Paul."
                gloss="N0=Nhum N0_V_N1">
                    <syntacticArgument id="a0" canonicalArgument="N0">
                        <syntacticFunction>subject</syntacticFunction>
                        <syntacticCategory>nounPhrase</syntacticCategory>
                        <semanticRestriction>human</semanticRestriction>
                        <semanticRestriction>intentional</semanticRestriction>
                    </syntacticArgument>
                    <syntacticArgument id="a1" canonicalArgument="N1">
                        <syntacticFunction>directObject</syntacticFunction>
                        <syntacticCategory>nounPhrase</syntacticCategory>
                        <semanticRestriction>human</semanticRestriction>
                    </syntacticArgument>
                </syntacticConstruction>
            </constructionSet>
       </sense>
</lexicalEntry>
```