Automatic Selectional Preference Acquisition for Latin verbs

Barbara McGillivray
University of Pisa
Italy
b.mcgillivray@ling.unipi.it

Abstract
We present a system that automatically induces Selectional Preferences (SPs) for Latin verbs from two treebanks by using Latin WordNet. Our method overcomes some of the problems connected with data sparseness and the small size of the input corpora. We also suggest a way to evaluate the acquired SPs on unseen events extracted from other Latin corpora.

1 Introduction
Automatic acquisition of semantic information from corpora is a challenge for research on low-resourced languages, especially when semantically annotated corpora are not available. Latin is definitely a high-resourced language for what concerns the number of available texts and traditional lexical resources such as dictionaries. Nevertheless, it is a low-resourced language from a computational point of view (McGillivray et al., 2009).

As far as NLP tools for Latin are concerned, parsing experiments with machine learning techniques are ongoing (Bamman and Crane, 2008; Passarotti and Ruffolo, forthcoming), although more work is still needed in this direction, especially given the small size of the training data. As a matter of fact, only three syntactically annotated Latin corpora are available (and still in progress): the Latin Dependency Treebank (LDT, 53,000 tokens) for classical Latin (Bamman and Crane, 2006), the Index Thomisticus Treebank (IT-TB, 54,000 tokens) for Thomas Aquinas’s works (Passarotti, 2007), and the PROIEL treebank (approximately 100,000 tokens) for the Bible (Haug and Jøndal, 2008). In addition, a Latin version of WordNet – Latin WordNet (LWN; Minozzi, 2009) – is being compiled, consisting of around 10,000 lemmas inserted in the multilingual structure of MultiWordNet (Bentivogli et al., 2004). The number and the size of these resources are small when compared with the corpora and the lexicons for modern languages, e. g. English.

Concerning semantic processing, no semantically annotated Latin corpus is available yet; building such a corpus manually would take considerable time and energy. Hence, research in computational semantics for Latin would benefit from exploiting the existing resources and tools through automatic lexical acquisition methods.

In this paper we deal with automatic acquisition of verbal selectional preferences (SPs) for Latin, i.e. the semantic preferences of verbs on their arguments: e. g. we expect the object position of the verb *edo* ‘eat’ to be mostly filled by nouns from the food domain. For this task, we propose a method inspired by Alishahi (2008) and outlined in an earlier version on the IT-TB in McGillivray (2009).

SPs are defined as probability distributions over semantic features extracted as sets of LWN nodes. The input data are two subcategorization lexicons automatically extracted from the LDT and the IT-TB (McGillivray and Passarotti, 2009).

Our main contribution is to create a new tool for semantic processing of Latin by adapting computational techniques developed for extant languages to the special case of Latin. A successful adaptation is contingent on overcoming corpus size differences. The way our model combines the syntactic information contained in the treebanks with the lexical semantic knowledge from LWN allows us to overcome some of the difficulties related to the small size of the input corpora. This is the main difference from corpora for modern languages, together with the absence of semantic annotation. Moreover, we face the problem of evaluating our system’s ability to generalize over unseen cases by using text occurrences, as access to human linguistic judgements is denied for Latin.

In the rest of the paper we will briefly summarize previous work on SP acquisition and motivate
our approach (section 2); we will then describe our system (section 3), report on first results and evaluation (section 4), and finally conclude by suggesting future directions of research (section 5).

2 Background and motivation

The state-of-the-art systems for automatic acquisition of verbal SPs collect argument headwords from a corpus (for example, apple, meat, salad as objects of eat) and then generalize the observed behaviour over unseen cases, either in the form of words (how likely is it to find sausage in the object position of eat?) or word classes (how likely is it to find VEGETABLE, FOOD, etc?).

WN-based approaches translate the generalization problem into estimating preference probabilities over a noun hierarchy and solve it by means of different statistical tools that use the input data as a training set: cf. inter al. Resnik (1993), Li and Abe (1998), Clark and Weir (1999). Agirre and Martinez (2001) acquire SPs for verb classes instead of single verb lemmas by using a semantically annotated corpus and WN.

Distributional methods aim at automatically inducing semantic classes from distributional data in corpora by means of various similarity measures and unsupervised clustering algorithms: cf. e. g. Rooth et al. (1999) and Erk (2007). Bamman and Crane (2008) is the only distributional approach dealing with Latin. They use an automatically parsed corpus of 3.5 million words, then calculate SPs with the log-likelihood test, and obtain an association score for each (verb, noun) pair.

The main difference between these previous systems and our case is the size of the input corpus. In fact, our dataset consists of subcategorization frames extracted from two relatively small treebanks, amounting to a little over 100,000 word tokens overall. This results in a large number of low-frequency (verb, noun) associations, which may not reflect the actual distributions of Latin verbs. This state improves if we group the observations into clusters. Such a method, proposed by Alishahi (2008), proved effective in our case.

The originality of this approach is an incremental clustering algorithm for verb occurrences called frames which are identified by specific syntactic and semantic features, such as the number of verbal arguments, the syntactic pattern, and the semantic properties of each argument, i. e. the WN hyponyms of the argument’s fillers. Based on a probabilistic measure of similarity between the frames’ features, the clustering produces larger sets called constructions. The constructions for a verb contribute to the next step, which acquires the verb’s SPs as semantic profiles, i. e. probability distributions over the semantic properties. The model exploits the structure of WN so that predictions over unseen cases are possible.

3 The model

The input data are two corpus-driven subcategorization lexicons which record the subcategorization frames of each verbal token occurring in the corpora: these frames contain morphosyntactic information on the verb’s arguments, as well as their lexical fillers. For example, ‘eo + A\text{\textunderscore\text{\textunderscore}Obj\text{\textunderscore\text{\textunderscore}acc}}\{\text{\textunderscore\text{\textunderscore}exsilium}\}’ represents an active occurrence of the verb eo ‘go’ with a prepositional phrase introduced by the preposition in ‘to, into’ and composed by an accusative noun phrase filled by the lemma exsilium ‘exile’, as in the sentence\footnote{Cicero, \textit{In Catilinam}, II, 7.}

\begin{equation}
\text{eat in exsilium}
\begin{align*}
go:\text{SBJV.PRS.3SG to exile:ACC.N.SG}
\end{align*}
\end{equation}

We illustrate how we adapted Alishahi’s definitions of frame features and formulae to our case. Alishahi uses a semantically annotated English corpus, so she defines the verb’s semantic primitives, the arguments’ participant roles and their semantic categories; since we do not have such annotation, we used the WN semantic information.

The syntactic feature of a frame ($ft_1$) is the set of syntactic slots of its verb’s subcategorization pattern, extracted from the lexicons. In the above example, ‘A\text{\textunderscore\text{\textunderscore}in\text{\textunderscore\textunderscore}Obj\text{\textunderscore\text{\textunderscore}acc}}’. In addition, the first type of semantic features of a frame ($ft_2$) collects the semantic properties of the verb’s arguments as the set of LWN synonyms and hyponyms of their fillers. In the previous example this is $\{\text{\textunderscore\text{\textunderscore}exsilium \ 'exile', proscriptio \ 'proscription', rejection, actio, actus \ 'act'}\}$\footnote{We listed the LWN node of the lemma exsilium, followed by its hyponyms; each node – apart from rejection, which is English and is not filled by a Latin lemma in LWN – is translated by the corresponding node in the English WN.} The second type of semantic features of a frame ($ft_3$) collects the semantic properties of the verb in the form of the verb’s synsets. In the above example, these are all synsets of eo ‘go’, among which ‘\{eo, gradior, grassor, ingredior, procedo, prodeo,
vado\} {\{progress, come\_on, come\_along, advance, get\_on, get\_along, shape\_up\}} in the English WN).

3.1 Clustering of frames

The constructions are incrementally built as new frames are included in them; a new frame $F$ is assigned to a construction $K$ if $F$ probabilistically shares some features with the frames in $K$ so that

$$K = \arg \max_k P(k|F) = \arg \max_k P(k)P(F|k),$$

where $k$ ranges over the set of all constructions, including the baseline $k_0 = \{F\}$. The prior probability $P(k)$ is calculated from the number of frames contained in $k$ divided by the total number of frames. Assuming that the frame features are independent, the posterior probability $P(F|k)$ is the product of three probabilities, each one corresponding to the probability that a feature displays in $k$ the same value it displays in $F$: $P_i(ft_i(F)|k)$ for $i = 1, 2, 3$:

$$P(F|k) = \prod_{i=1,2,3} P_i(ft_i(F)|k)$$

We estimated the probability of a match between the value of $ft_1$ in $k$ and the value of $ft_1$ in $F$ as the sum of the syntactic scores between $F$ and each frame $h$ contained in $k$, divided by the number $n_k$ of frames in $k$:

$$P(ft_1(F)|k) = \frac{\sum_{h \in k} \text{sytscore}(h,F)}{n_k}$$

where the syntactic score $\text{sytscore}(h,F) = \frac{|\text{SCS}(h) \cap \text{SCS}(F)|}{|\text{SCS}(F)|}$ calculates the number of syntactic slots shared by $h$ and $F$ over the number of slots in $F$. $P(ft_1(F)|k)$ is 1 when all the frames in $k$ contain all the syntactic slots of $F$.

For each argument position $a$, we estimated the probability $P(ft_2(F)|k)$ as the sum of the semantic scores between $F$ and each frame $h$ in $k$:

$$P(ft_2(F)|k) = \frac{\sum_{h \in k} \text{semscore}(h,F)}{n_k}$$

where the semantic score $\text{semscore}(h,F) = \frac{|\text{S}(h) \cap \text{S}(F)|}{|\text{S}(F)|}$ counts the overlap between the semantic properties $S(h)$ of $h$ (i.e. the LWN hypernyms of the fillers in $h$) and the semantic properties $S(F)$ of $F$ (for argument $a$), over $|\text{S}(F)|$.

$$P(ft_3(F)|k) = \frac{\sum_{h \in k} \text{syns\_score}(h,F)}{n_k}$$

where the synset score $\text{syns\_score}(h,F) = \frac{|\text{Synsets}(\text{node}(h))\cap \text{Synsets}(\text{node}(F))|}{|\text{Synsets}(\text{node}(F))|}$ calculates the overlap between the synsets for the verb in $h$ and the synsets for the verb in $F$ over the number of synsets for the verb in $F$.

We introduced the syntactic and synset scores in order to account for a frequent phenomenon in our data: the partial matches between the values of the features in $F$ and in $k$.

3.2 Selectional preferences

The clustering algorithm defines the set of constructions in which the generalization step over unseen cases is performed. SPs are defined as semantic profiles, that is, probability distributions over the semantic properties, i.e. LWN nodes. For example, we get the probability of the node ‘act’ in the position ‘A\_in(Obj[ace])’ for eo ‘go’.

If $s$ is a semantic property and $a$ an argument position for a verb $v$, the semantic profile $P_a(s|v)$ is the sum of $P_a(s,k|v)$ over all constructions $k$ containing $v$ or a WN-synonym of $v$, i.e. a verb contained in one or more synsets for $v$. $P_a(s,k|v)$ is approximated as $\frac{P(k,v)P_a(s|v)}{P(v)}$, where $P(k,v)$ is estimated as $\sum_{k} n_k \cdot \frac{\text{freq}(k,v)}{\text{freq}(v)}$.

To estimate $P_a(s,k|v)$ we consider each frame $h$ in $k$ and account for: a) the similarity between $v$ and the verb in $h$; b) the similarity between $s$ and the fillers of $h$. This is achieved by calculating a similarity score between $h$, $v$, $a$, and $s$, defined as:

$$\text{syns\_score}(v,H) \cdot \frac{\sum_f |S(f) \cap S_h|}{N_{fil}(h,a)} \tag{1}$$

where $V(h)$ in (1) contains the verbs of $h$, $N_{fil}(h,a)$ counts the $a$-fillers in $h$, $f$ ranges in the set of $a$-fillers in $h$, $S(f)$ contains the semantic properties for $f$ and $|S(f) \cap S_h|$ is 1 when $s$ appears in $S(f)$ and 0 otherwise.

$P_a(s,k|v)$ is thus obtained by normalizing the sum of these similarity scores over all frames in $k$, divided by the total number of frames in $k$ containing $v$ or its synonyms.

The similarity scores weight the contributions of the synonyms of $v$, whose fillers play a role in the generalization step. This is our innovation with respect to Alishahi (2008)'s system. It was introduced because of the sparseness of our data, where

\footnote{The algorithm uses smoothed versions of all the previous formulae by adding a very small constant so that the probabilities are never 0.}
4 Results and evaluation

The clustering algorithm was run on 15509 frames and it generated 7105 constructions. Table 1 displays the 5 constructions assigned to the 9 frames where the verb *introduco* ‘bring in, introduce’ occurs. Note the semantic similarity between *addo* ‘add to, bring to’, *inmitto* ‘send against, insert’, *induco* ‘bring forward, introduce’ and *introduco*, and the similarity between the syntactic patterns and the argument fillers within the same construction. For example, *finis* ‘end, borders’ and *effectus* ‘result’ share the semantic properties ATTRIBUTE, COGNITIO ‘cognition’, CONSCIENTIA ‘conscience’, EVENTUM ‘event’, among others.

The vast majority of constructions contain less than 4 frames. This contrasts with the more general constructions found by Alishahi (2008) and can be explained by several factors. First, the coverage of LWN is quite low with respect to the fillers in our dataset. In fact, 782 fillers out of 2408 could not be assigned to any LWN synset; for these lemmas the semantic scores with all the other nouns are 0, causing probabilities lower than the baseline; this results in assigning the frame to the singleton construction consisting of the frame itself. The same happens for fillers consisting of verbal lemmas, participles, pronouns and named entities, which amount to a third of the total number. Furthermore, the data are not tagged by sense and the system deals with noun ambiguity by listing together all synsets of a word *n* (and their hyponyms) to form the semantic properties for *n*: consequently, each sense contributes to the semantic description of *n* in relation to the number of hyponyms it carries, rather than to its observed frequency. Finally, a common problem in SP acquisition systems is the noise in the data, including tagging and metaphorical usages. This problem is even greater in our case, where the small size of the data underestimates the variance and therefore overestimates the contribution of noisy observations. Metaphorical and abstract usages are especially frequent in the data from the IT-TB, due to the philosophical domain of the texts.

As to the SP acquisition, we ran the system on all constructions generated by the clustering. We excluded the pronouns occurring as argument fillers, and manually tagged the named entities. For each verb lemma and slot we obtained a probability distribution over the 6608 LWN noun nodes.

Table 2 displays the 20 semantic properties with the highest SP probabilities as ablative arguments of *ascendo* ‘ascend’ introduced by *de* ‘down from’, ‘out of’. This semantic profile was created from the following fillers for the verbs contained in the constructions for *ascendo* and its synonyms: abyssus ‘abyss’, fiumus ‘smoke’, lacus ‘lake’, machina ‘machine’, manus ‘hand’, negotiatio ‘business’, mare ‘sea’, os ‘mouth’, templum ‘temple’, terra ‘land’. These nouns are well represented by the semantic properties related to water and physical places. Note also the high rank of general properties like actio ‘act’, which are associated to a large number of fillers and thus generally get a high probability.

Regarding evaluation, we are interested in testing two properties of our model: calibration and discrimination. Calibration is related to the model’s ability to distinguish between high and low probabilities. We verify that our model is
adequately calibrated, since its SP distribution is always very skewed (cf. figure 1). Therefore, the model is able to assign a high probability to a small set of nouns (preferred nouns) and a low probability to a large set of nouns (the rest), thus performing better than the baseline model, defined as the model that assigns the uniform distribution over all nouns (4724 LWN leaf nodes). Moreover, our model’s entropy is always lower than the baseline: 12.2 vs. the 6.9-11.3 range; by the maximum entropy principle, this confirms that the system uses some information for estimating the probabilities: LWN structure, co-occurrence frequency, syntactic patterns. However, we have no guarantee that the model uses this information sensibly. For this, we test the system’s discrimination potential, i.e. its ability to correctly estimate the SP probability of each single LWN node.

Table 3: 15 nouns with the highest probabilities as accusative objects of *dico* ‘say’.

| noun          | SP probability |
|---------------|----------------|
| pars ‘part’   | 0.0029         |
| locus ‘place’ | 0.0026         |
| ratio ‘account’ ‘reason’, ‘opinion’ | 0.0023 |
| respectus ‘consideration’ | 0.0022 |
| caput ‘head’  | 0.0022         |
| anima ‘soul’  | 0.0021         |
| animus ‘soul’, ‘spirit’ | 0.0020 |
| figura ‘form’, ‘figure’ | 0.0020 |
| spiritus ‘spirit’ | 0.0020 |
| corpus ‘body’ | 0.0019         |
| sententia ‘judgement’ | 0.0019 |
| finitio ‘limit’, ‘definition’ | 0.0019 |
| species ‘sight’, ‘appearance’ | 0.0019 |

For what concerns evaluating the SP probability assigned to nouns unseen in the training set, Alishahi (2008) follows the approach suggested by Resnik (1993), using human plausibility judgements on verb-noun pairs. Given the absence of native speakers of Latin, we used random occurrences in corpora, considered as positive examples of plausible argument fillers; on the other hand, we cannot extract non-plausible fillers from a corpus unless we use a frequency-based criterion. However, we can measure how well our system predicts the probability of these unseen events.

As a preliminary evaluation experiment, we randomly selected from our corpora a list of 19 high-frequency verbs (freq.>51) and 7 medium-frequency verbs (11<freq.<50), for each of which we chose an interesting argument slot. Then we randomly extracted one filler for each such pair from two collections of Latin texts (*Perseus Digital Library* and *Corpus Thomisticum*), provided that it was not in the training set. The semantic score in equation 1 on page 3 is then calculated between the set of semantic properties of *n* and that for *f*, to obtain the probability of finding the random filler *n* as an argument for a verb *v*.

For each of the 26 (verb, slot) pairs, we looked at three measures of central tendency: mean, median and the value of the third quartile, which were compared with the probability assigned by the model to the random filler. If this probability was higher than the measure, the outcome was considered a success. The successes were 22 for the mean, 25 for the median and 19 for the third quartile.4 For all three measures a binomial test found the success rate to be statistically significant at the 5% level. For example, table 3 and figure 1 show that the filler for *dico+A_Obj[acc]* in the evaluation set – *sententia* ‘judgement’ – is ranked 13th within the verb’s semantic profile.

5 Conclusion and future work

We proposed a method for automatically acquiring probabilistic SP for Latin verbs from a small corpus using the WN hierarchy; we suggested some
new strategies for tackling the data sparseness in
the crucial generalization step over unseen cases.
Our work also contributes to the state of the art in
semantic processing of Latin by integrating syn-
tactic information from annotated corpora with the
lexical resource LWN. This demonstrates the use-
fulness of the method for small corpora and the
relevance of computational approaches for histor-
ical linguistics.

In order to measure the impact of the frame
collectors for the SP acquisition, we plan to run the
system for SP acquisition without performing the
clustering step, thus defining all constructions as
singleton sets containing one frame each. Finally,
an extensive evaluation will require a more com-
prehensive set, composed of a higher number of
unseen argument fillers; from the frequencies of
these nouns, it will be possible to directly compare
plausible arguments (high frequency) and implaus-
able ones (low frequency). For this, a larger auto-
matically parsed corpus will be necessary.

6 Acknowledgements

We wish to thank Afra Alishahi, Stefano Minozzi
and three anonymous reviewers.

References

E. Agirre and D. Martinez. 2001. Learning class-to-
class selectional preferences. In Proceedings of the
ACL/EACL 2001 Workshop on Computational Nat-
ural Language Learning (CoNLL-2001), pages 1–8.

A. Alishahi. 2008. A probabilistic model of early arg-
ument structure acquisition. Ph.D. thesis, Depart-
ment of Computer Science, University of Toronto.

D. Bamman and G. Crane. 2006. The design and use
of a Latin dependency treebank. In Proceedings of
the Fifth International Workshop on Treebanks and
Linguistic Theories, pages 67–78. UFAL MFF UK.

D. Bamman and G. Crane. 2008. Building a dynamic
lexicon from a digital library. In Proceedings of the
8th ACM/IEEE-CS Joint Conference on Digital Li-
braries, pages 11–20.

L. Bentivogli, P. Forner, and and Pianta E. Magnini,
B. 2004. Revising wordnet domains hierarchy: Sem-
antics, coverage, and balancing. In Proceedings of
COLING Workshop on Multilingual Linguistic Re-
sources, pages 101–108.

S. Clark and D. Weir. 1999. An iterative approach
to estimating frequencies over a semantic hierarchy.
In Proceedings of the Joint SIGDAT Conference on
Empirical Methods in Natural Language Processing
and Very Large Corpora. University of Maryland,
pages 258–265.

K. Erk. 2007. A simple, similarity-based model for
selectional preferences. In Proceedings of the 45th
Annual Meeting of the Association for Computa-
tional Linguistics, pages 216–223.

D. T. T. Haug and M. L. Jondal. 2008. Creating a par-
allel treebank of the old Indo-European Bible trans-
lations. In Proceedings of Language Technologies
for Cultural Heritage Workshop, pages 27–34.

H. Li and N. Abe. 1998. Generalizing case frames
using a thesaurus and the MDL principle. Compu-
tational Linguistics, 24(2):217–244.

B. McGillivray and M. Passarotti. 2009. The devel-
opment of the Index Thomisticus Treebank Valency
Lexicon. In Proceedings of the Workshop on Lan-
guage Technology and Resources for Cultural Her-
itage, Social Sciences, Humanities, and Education,
pages 33–40.

B. McGillivray, M. Passarotti, and P. Ruffolo. 2009.
The Index Thomisticus treebank project: Annotation,
parsening and valency lexicon. TAL, 50(2):103–127.

B. McGillivray. 2009. Selectional Preferences from
a Latin treebank. In Przepiorkowski A. Passarotti,
M., S. Raynaud, and F. van Eynde, editors, Proceed-
ings of the Eighth International Workshop on Tree-
banks and Linguistic Theories (TLT8), pages 131–
136. EDUCatt.

S. Minozzi. 2009. The Latin Wordnet project. In P.
Anreiter and M. Kienpointner, editors, Proceed-
ings of the 15th International Colloquium on
Latin Linguistics (ICLL), Innsbrucker Beiträge zur
Sprachwissenschaft.

M. Passarotti and P. Ruffolo. forthcoming. Parsing the
Index Thomisticus Treebank, some preliminary
results. In P. Anreiter and M. Kienpointner, edi-
tors, Proceedings of the 15th International Collo-
quium on Latin Linguistics, Innsbrucker Beiträge zur
Sprachwissenschaft.

M. Passarotti. 2007. Verso il Lessico Tomistico Bi-
culturale. La treebank dell’Index Thomisticus. In R.
Petrilli and D. Femia, editors, Atti del XIII Con-
gresso Nazionale della Società di Filosofia del Lin-
guaggio, pages 187–205.

P. Resnik. 1993. Selection and Information: A Class-
Based Approach to Lexical Relationships. Ph.D.
thesis, University of Pennsylvania.

M. Rooth, S. Riezler, D. Prescher, G. Carroll, and
F. Beil. 1999. Inducing a semantically annotated
lexicon via EM-based clustering. In Proceedings of
the 37th Annual Meeting of the Association for Com-
putational Linguistics, pages 104–111.