

# Dottorato di Ricerca in Linguistica Generale, Storica, Applicata, Computazionale e delle Lingue Moderne <br> L-LIN/01 

## Tesi di dottorato

A computational approach to Latin verbs: new resources and methods

Candidato:
Dott. Barbara McGillivray

Tutori:
Dott. Alessandro Lenci Prof. Giovanna Marotta

Presidente: Prof. Giovanna Marotta

## Contents

Acknowledgements ..... ix
List of abbreviations ..... xi
List of figures ..... xiii
List of tables ..... xvii
1 Introduction ..... 1
1.1 Computational linguistics for Latin ..... 1
1.1.1 Computational approaches ..... 2
1.1.2 A less-resourced language ..... 5
1.2 Aims ..... 6
1.3 Scope. Corpora and language ..... 7
1.4 Overview of the thesis ..... 12
2 Background ..... 15
2.1 Introduction ..... 15
2.2 Arguments and adjuncts ..... 15
2.2.1 Definitions ..... 15
2.2.2 Approaches ..... 19
2.3 Latin verbal arguments ..... 27
2.4 Summary ..... 35
3 Computational resources and tools for Latin ..... 37
3.1 Introduction ..... 37
3.2 Corpora and Natural Language Processing tools ..... 37
3.2.1 Digital editions ..... 38
3.2.2 Morphological annotation of corpora ..... 39
3.2.3 Syntactic annotation of corpora ..... 39
3.3 Lexical databases ..... 42
3.3.1 Latin WordNet ..... 42
3.4 Summary ..... 42
4 The subcategorization lexicon ..... 43
4.1 Introduction ..... 43
4.2 Motivation ..... 44
4.3 Background ..... 44
4.3.1 Intuition-based and corpus-based dictionaries ..... 45
4.3.2 Manually annotated corpora ..... 48
4.3.3 Automatically acquired lexicons ..... 49
4.4 Our contribution ..... 53
4.4.1 A new resource for Latin ..... 53
4.4.2 Overview of the system ..... 54
4.5 Subcategorization extraction ..... 55
4.5.1 Dependency annotation for Latin ..... 55
4.5.2 The annotation of the LDT and the IT-TB ..... 57
4.5.3 Overview of the database queries ..... 67
4.6 Subcategorization representation ..... 68
4.6.1 Subcategorization frames and classes ..... 68
4.6.2 The entries of the lexicon ..... 69
4.7 Quantitative data ..... 70
4.7.1 Verb frequencies ..... 72
4.7.2 Subcategorization frequencies ..... 74
4.7.3 Word order patterns ..... 75
4.8 Remarks and future work ..... 77
4.8.1 Features of the lexicon ..... 77
4.8.2 Future directions of research ..... 79
4.9 Summary ..... 79
5 Selectional Preferences ..... 81
5.1 Introduction ..... 81
5.2 Previous work ..... 82
5.2.1 Knowledge-based approaches ..... 82
5.2.2 Knowledge-free approaches ..... 87
5.3 Our model ..... 91
5.3.1 Overview ..... 96
5.3.2 Selectional preference acquisition ..... 100
5.4 Evaluation ..... 109
5.4.1 The probability distribution ..... 109
5.4.2 The evaluation set ..... 111
5.4.3 A distributional measure of semantic similarity ..... 114
5.5 Conclusions and future work ..... 117
6 Introduction to Latin preverbs ..... 119
6.1 Introduction ..... 119
6.2 A first definition of preverbs ..... 119
6.3 Previous work on preverbs ..... 120
6.3.1 Overview of the literature ..... 120
6.3.2 Origin and development of Latin preverbs ..... 122
6.3.3 Spatial preverbs and motion events ..... 127
6.4 Argument structure of Latin preverbed verbs ..... 130
6.4.1 Synchronic overview ..... 131
6.4.2 Diachronic hypotheses ..... 134
7 A corpus-based study on Latin spatial preverbs ..... 139
7.1 Introduction ..... 139
7.2 Research design ..... 139
7.3 The corpora and the dataset ..... 141
7.3.1 Corpora ..... 141
7.3.2 Dataset ..... 142
7.4 First investigation ..... 144
7.4.1 Dataset ..... 145
7.4.2 Statistical tests ..... 146
7.4.3 A multivariate approach ..... 152
7.4.4 First conclusions ..... 154
7.5 Second investigation ..... 154
7.5.1 The methodology ..... 154
7.5.2 Relevant variables ..... 155
7.6 Third investigation ..... 160
7.6.1 Dataset ..... 160
7.6.2 Analysis ..... 161
7.7 Summary and conclusions ..... 163
7.7.1 Future work ..... 164
8 Conclusions ..... 165
8.1 Introduction ..... 165
8.2 Aims and contributions ..... 165
8.2.1 New resource ..... 165
8.2.2 New methods ..... 166
8.3 Future research ..... 167
8.4 General conclusions ..... 168
Appendices ..... 169
A Background to chapter 4 ..... 171
A. 1 Annotation format of the treebanks ..... 171
A. 2 Database queries for subcategorization extraction ..... 173
A.2.1 Searching for verbal tokens ..... 174
A.2.2 Searching for verbal arguments: direct head-argument dependencies1 ..... 174
A.2.3 Searching for verbal arguments. Indirect head-argument dependencies ..... 176
A.2.4 Searching for verbal arguments. Some difficult cases ..... 180
A. 3 Subcategorization frames and classes ..... 181
A. 4 Frequency information in the lexicon ..... 184
A.4.1 How to extract frequency information ..... 184
A.4.2 SC frequency and verb-SC frequency ..... 184
B Statistical background to chapter 5 ..... 193
B. 1 Bayesian statistics ..... 193
B.1.1 Prior probability and posterior probability ..... 193
B.1.2 Bayes' Theorem ..... 194
B. 2 Entropy ..... 194
B. 3 Knowledge-based approaches in SP acquisition ..... 195
B.3.1 Resnik (1993) ..... 195
B.3.2 Ribas (1995) ..... 197
B.3.3 Li and Abe (1998) ..... 197
B.3.4 McCarthy (2001) ..... 198
B.3.5 Clark and Weir (1999) ..... 199
B.3.6 Abney and Light (1999) ..... 199
B.3.7 Ciaramita and Johnson (2000) ..... 199
B.3.8 Agirre and Martinez (2001) ..... 200
B. 4 Vector spaces ..... 201
B.4.1 Data format ..... 201
B.4.2 Vectors ..... 201
B. 5 Clustering ..... 202
B.5.1 Overview ..... 203
B.5.2 Hard agglomerative hierarchical clustering ..... 205
B. 6 Knowledge-free approaches in SP acquisition ..... 207
B.6.1 Rooth et al. (1999a) ..... 207
B.6.2 Erk (2007) ..... 208
B.6.3 Bamman and Crane (2008) ..... 208
B. 7 Data from Latin WordNet ..... 209
B.7. 1 Synsets ..... 209
B.7.2 Examples of semantic properties ..... 214
B. 8 Formulae for the probabilistic model ..... 215
B.8.1 Clustering of frames ..... 215
B.8.2 SP probability distribution ..... 216
B. 9 Formulae for the evaluation of the SP system ..... 217
B.9.1 Corpus frequencies ..... 223
B.9.2 Distributional similarity ..... 26
C Statistical background to chapter 7 ..... 231
C. 1 Introduction ..... 231
C. 2 Pearson's $\chi^{2}$ test of independence ..... 231
C. 3 Linear regression models ..... 234
C.3.1 Linear functions ..... 234
C.3.2 Linear regression ..... 235
C.3.3 Generalized Linear Models ..... 239
C.3.4 Mixed effect models ..... 239
C.3.5 The model for the first investigation ..... 239
C. 4 Correspondence Analysis ..... 243
C.4.1 Introduction ..... 243
C.4.2 The idea ..... 244
C.4.3 Procedure ..... 244
C.4.4 Multiple Correspondence Analysis ..... 251
C.4.5 The second investigation ..... 251
C.4.6 The third investigation ..... 255
C. 5 Configural Frequency Analysis ..... 25
C.5.1 First attempts ..... 258
C.5.2 Configural Frequency Analysis ..... 259
C.5.3 A Configural Frequency Analysis for Direct Object Marking ..... 259
CONTENTS ..... vii
Bibliography ..... 261
Primary sources ..... 263
Dictionaries and Thesauri ..... 263
References ..... 264
Index ..... 285

## Acknowledgements

Firstly, I wish to thank my supervisors, Alessandro Lenci and Giovanna Marotta, for their guidance and inspiration throughout these years. By believing in this project, they have encouraged me with helpful comments and the contribution of their valuable and broad knowledge.

I am also grateful to a number of scholars who have given me their precious advice during fuitful collaborations: Daniel Apollon, Eystein Dahl, Chiara Fedriani, Aldo Frigerio, Gard Jenset, Christer Johansson, Simonetta Montemagni, Marco Passarotti, Vito Pirrelli, Savina Raynaud, and Paolo Ruffolo. I thank Afra Alishahi for kindly discussing some details of her research with me.

I would like to express all my gratitude to my mother for always believing in me and being on my side. Finally, special thanks to Gard, who has turned these difficult years into a wonderful experience.

## List of abbreviations

BBN Bayesian Belief Network<br>BNC British National Corpus<br>CA Correspondence Analysis<br>CFA Configural Frequency Analysis<br>DAG Directed Acyclic Graph<br>DG Dependency Grammar<br>DOM Direct Object Marking<br>EM Expectation Maximization<br>FGD Functional Generative Description<br>GLM Generalized Linear Model<br>HMM Hidden Markov Model<br>IQR Interquartile Range<br>IT Index Thomisticus<br>IT-TB Index Thomisticus Treebank<br>LDT Latin Dependency Treebank<br>LWN Latin WordNet<br>MAD Median Absolute Deviation<br>MCA Multiple Correpondence Analysis<br>MDL Minimum Description Length<br>MLE Maximum Likelihood Estimate<br>MWN MultiWordNet<br>NP Noun Phrase<br>PDT Prague Dependency Treebank<br>PoS Part of Speech<br>PP Prepositional Phrase<br>PSG Phrase Structure Grammar<br>PV Preverbed Verb<br>PWN Princeton WordNet<br>SC Subcategorization<br>SCC Subcategorization Class<br>SCF Subcategorization Frame<br>SCS Subcategorization Structure<br>SP Selectional Preference<br>TLL Thesaurus Linguae Latinae<br>VP Verbal Phrase<br>WN WordNet<br>WSD Word Sense Disambiguation

## List of Figures

2.1 A dependency representation of sentence (14) ..... 20
2.2 PSG tree (left) and DG tree (right) for sentence (14) on page 20. ..... 22
2.3 X-bar tree for sentence (14) on page 20. ..... 24
4.1 Valency frames for the Czech verb dát 'give', the noun dopis 'a letter' and the adjective plny 'full' in the FGD formalism. ..... 47
4.2 A DG representation of sentence (2) ..... 57
4.3 The dependency tree of sentence (3) from the IT-TB. ..... 58
4.4 The dependency tree of sentence (4) from the LDT. ..... 60
4.5 The dependency tree of sentence (5) from the LDT. ..... 61
4.6 The dependency tree of sentence (6) from the LDT. ..... 62
4.7 The dependency tree of sentence (7) from the LDT. ..... 63
4.8 The dependency tree of sentence (8) from the LDT. ..... 63
4.9 The dependency tree of sentence (9) from the LDT. ..... 64
4.10 Tree of sentence (10). ..... 66
4.11 Sentence (2) (page 57) of the LDT as it is loaded into the database table for the LDT. ..... 67
4.12 Lexical entry for the verb impono from the SC lexicon for the LDT. For the definition of $S C C_{2}$, see section A.3. ..... 70
4.13 Lexical entry for the verb impono from the SC lexicon for the IT-TB ..... 71
4.14 LDT sublexicon ..... 72
4.15 IT-TB sublexicon ..... 72
4.16 Distribution of verb frequencies in the two sublexicons. ..... 72
5.1 Example portion of a WN-like hierarchy for some possible subjects of fly. ..... 83
5.2 Two-dimensional plot from table 5.2, representing the subjects of fly and buzz (absolute frequencies) ..... 89
5.3 Two-dimensional plot from table 5.3, representing the subjects of fly and buzz (relative frequencies). ..... 89
5.4 Dendrogram from a clustering for the data in table 5.2 on page 88 . ..... 90
5.5 Portion of the LWN noun hierarchy containing the nodes 'terra/11' (the 11th synset for terra) and 'fons/3' (the third synset for fons). ..... 93
5.6 Portion of the LWN noun hierarchy containing the fillers of cado in our corpus (except for terra 'earth'): culpa 'fault', esse 'being', fons 'source', forma 'form' and pars 'part'. ..... 95
5.7 Portion of the LWN noun hierarchy containing the fillers of descendo in our corpus: homo 'man', solium 'throne' and terra 'earth'. ..... 95
5.8 Portion of the LWN noun hierarchy containing the semantic properties of grando 'hail'. ..... 98
5.9 Decreasing SP probabilities of the LWN leaf nodes for the objects of dico 'say'. ..... 110
5.10 Mean, median and value of the third quantile for the objects of dico 'say'. ..... 112
7.1 Barplot of the distribution of 'CasePrev' (light grey bars) and 'Preposi- tion' (dark grey bars) by author; the percentage over all verbal tokens by author is displayed. ..... 147
7.2 The variables 'author' and 'construction' plotted against each other in the first dataset (absolute frequencies) ..... 147
7.3 The variables 'author' and 'construction' plotted against each other in the first dataset. The proportions of the two values of 'construction' in each author are displayed ..... 148
7.4 Association plot for the $\chi^{2}$ statistics on the two values for the variable 'construction' by author in the first dataset. ..... 149
7.5 Association plot for the $\chi^{2}$ statistics on the two values for the variable 'construction' by author in the first dataset. We merged the data for Caesar, Cicero and Sallust under the label 'late republic', and the data for Ovid, Propertius and Vergil under the label 'poets'. ..... 151
7.6 Plot from MCA on the variables 'construction' and 'author'. The first axis accounts for $66.5 \%$ of the explained inertia, the second axis for $22.1 \%$. The labels 'construction.2CaseNonPrev' and 'Ovid' overlap. ..... 156
7.7 Plot of the MCA for the second investigation on the interaction between 'construction', 'era', 'preverb', 'animacy', and 'class'. Percentages of explained inertia: 27.6 (first axis), 20.5 (second axis) ..... 158
7.8 Plot of the MCA for the third investigation. The first axis accounts for $79.5 \%$ of the explained inertia, while the second axis accounts for $3.2 \%$. ..... 162
A. 1 Sentence (3) (page 58) of the IT-TB as it is loaded into the database. ..... 173
A. 2 A 'direct' dependency between a verbal head (annotated with any syn- tactic tag) and an argument node (tagged as 'Sb', 'Obj', 'Pnom', or 'OComp'). ..... 175
A. 3 An indirect dependency between a verbal head and its argument via an 'AuxP' node. This example is taken from sentence (4). ..... 176
A. 4 An indirect dependency between a verbal head and its arguments via a 'Coord' node. This example is taken from sentence (4) ..... 177
A. 5 Coordinated PPs with one 'AuxP' node. ..... 178
A. 6 Dependency tree of sentence (1) from the IT-TB. ..... 178
A. 7 Coordinated PPs with two 'AuxP' nodes. ..... 178
A. 8 "External" argument nodes. ..... 180
A. 9 Dependency tree of sentence (3) from the IT-TB ..... 181
A. 10 Dependency tree of sentence (4) ..... 183
A. 11 Verb frequencies and number of SCSs in the LDT lexicon. See table A. 12 for the ten most frequent verbs. ..... 189
A. 12 Verb frequencies and number of SCSs in the IT-TB lexicon. See table A. 12 for the ten most frequent verbs. ..... 190
B. 1 Example portion of a WN-like hierarchy for some possible subjects of fly listed in table B.1. ..... 195
B. 2 2-dimensional Cartesian space displaying the points corresponding to the rows of table B. 4 ..... 202
B. 3 Unit circle, the two vectors $(4,0)$ and $(2,1)$, the angle $\alpha$ between them and its cosine. ..... 204
B. 4 Dendrogram from the clustering represented in table B. 9 ..... 207
B. 5 SP probability distribution for the objects of dico 'say'. ..... 219
C. 1 Example of the line with equation $y=2 x+1$ and intercept $b=1$. The point $P=(1.5,4)$ is represented. ..... 234
C. 2 Data points from table C. 3 relative to speed of cars and their distance to stop. ..... 237
C. 3 Regression line of the data points plotted in figure C.2 ..... 238
C. 4 Binned plot of the residuals of the linear model against fitted. ..... 241
C. 5 Geometrical visualisation of the row profiles from table C. 6 including the average profile. ..... 246
C. 6 Four points in a two-dimensional space. ..... 248
C. 7 One-dimensional projection from figure C.6. ..... 249
C. 8 CA plot for the table C. 5 ..... 250
C. 9 Plot of the MCA for the second investigation on the interaction between 'construction' and 'era'. ..... 252
C. 10 First axis of the MCA for the second investigation on the interaction between 'construction' and 'author'. ..... 253
C. 11 Second axis of the MCA for the second investigation on the interaction between 'construction' and 'author'. ..... 254
C. 12 Plot of the MCA for the second investigation on the interaction between 'construction', 'era', 'preverb', 'animacy', and 'class'. ..... 256
C. 13 Plot of the MCA for the third investigation. ..... 256

## List of Tables

1.1 Example of annotation of the Latin sentence sese iactabit audacia "au- dacity will throw itself" ..... 4
1.2 Composition of the reference corpora for our study by author: dates of birth and death of the author, literary age, genre and number of word tokens are provided. ..... 8
4.1 Five most frequent SCFs for go in Valex. ..... 52
4.2 SCFs for aller 'go' in LexSchem. The first column contains the SCF, the second the absolute frequency and the third the relative frequency. ..... 53
4.3 Sentences obtained from sentence (1) on page 55 by changing the order of the words, and corresponding DG representations ..... 56
4.4 Non-projectivity in the LDT and in the IT-TB. ..... 57
4.5 Complete Latin tagset adopted by the LDT and the IT-TB. ..... 59
4.6 Number of verbal tokens $(|T|)$ and verbal types (forms $\left|V_{T f}\right|$ and lemmas $\left.\left|V_{T l}\right|\right)$ in the LDT and in the IT-TB sublexicons. $\frac{\left|V_{T f}\right|}{|T|}$ and $\frac{\left|V_{T l}\right|}{|T|}$ show the type/token ratios. ..... 72
4.7 Six most frequent verbs in the LDT and in the IT-TB. The lemma dico refers to the verb dico, -is 'say' (third conjugation class) and not to dico, -as 'dedicate' (first conjugation class) ..... 73
4.8 Counts of verbal tokens (lemmas), zero-valency usages, active and passive verbal forms in the LDT and in the IT-TB sublexicons. ..... 74
4.9 Counts of the different SCSs in the LDT and in the IT-TB sublexi- cons. More precisely, the SCF and SCC structures referred to here are $S C F_{1_{\text {voice }}}$ and $S C C_{2_{\text {voice }}}$, respectively; for the definition of $S C F_{1}$ and $S C C_{2}$, see section A.3. ..... 74
4.1015 most frequent types of $S C C_{\text {voice-morph }}$ and their frequencies in the LDT and IT-TB lexicons. More precisely, the SCC structures referred to here are $S C C_{2_{\text {voice-morph }}}$; for the definition of $S C C_{2}$, see section A.3. . . ..... 75
4.11 Most frequent functional labels (syntactic functions) and lemmas of the intermediate nodes between the verbs and their arguments. ..... 76
4.12 Frequencies of the word order patterns ' VO ' and ' OV ' in the two sublex- icons. ..... 76
4.13 Frequencies of word order patterns in the LDT lexicon without the data for Jerome, contrasted with the IT-TB sublexicon plus the data for Jerome ..... 77
5.1 Example corpus frequencies for the subjects of $f l y$. ..... 88
5.2 Constructed corpus frequencies of the nouns in table 5.1 as subjects of fly and buzz. ..... 88
5.3 Relative frequencies of the nouns in table 5.1 as subjects of $f l y$ and buzz. ..... 90
5.4 Synset number 11 for terra 'earth' and synset number 3 for fons 'source' in LWN, with their corresponding English synsets. ..... 93
5.5 The frame for sentence (2). ..... 97
5.6 Spelling differences between LWN synsets and lemmas in LDT and in IT-TB. The last column reports the standard we chose. For the LWN lemmas, we indicate the synset number after the '/' sign. ..... 99
5.7 Synset for descendo 'come down' in our enriched version of LWN, with its corresponding English synset in PWN. ..... 99
5.8 Selected synsets for porta, entrata, door and entrance in MWN, with their domain and glosses. ..... 101
5.9 Synsets for porta, entrata, door and entrance in MWN ..... 102
5.10 Schema of a clustering algorithm. ..... 103
5.11 Synsets of cado 'fall' and descendo 'descend'. For each sense of these two verbs, the first row contains the LWN synset, while the second row contains their corresponding English synsets in MWN. ..... 105
5.12 Constructions for the frames containing the verb introduco 'bring in'. ..... 106
5.13 Different categories of fillers in our lexicon and their frequency. ..... 106
5.14 Constructions for the frames containing the synonyms of the verb intro- duco 'insert, bring in' in LWN. ..... 108
5.1515 nouns with the highest probabilities as accusative objects of dico 'say'. 1 ..... 110
5.16 Vectors of the verbs voco 'call' and nomino 'name' in a multidimensional space defined by the SC structures with which these two verbs occur in the two different 3 -million-token parsed corpora. See table 5.18 for the description of the SC structures $S_{1}$, etc. ..... 115
5.17 Vectors of the nouns terra 'earth' and caelum 'sky' in a multidimensional space defined by the first (verb, slot) pairs with which these nouns occur in the two different 3 -million-token parsed corpora. See table 5.19 for the description of the structures $S_{1}$, etc ..... 115
5.18 Description of the SC structures in table 5.16. ..... 116
5.19 Description of the structures in table 5.17. ..... 117
6.1 Grammaticalization chain for the origin of preverbs in Latin. ..... 125
7.1 Number of word tokens, verb tokens and PV tokens (in bold) by author in the data. The number of verb tokens for Plautus is not recorded because the original data are not tagged by part of speech. ..... 145
7.2 Frequency counts of 'CasePrev' and 'Preposition' by author. ..... 146
7.3 Expected frequencies corresponding to the frequencies in table 7.2. ..... 150
A. 1 XML code for sentence (2) (page 57) in the LDT. ..... 171
A. 2 CSTS code for sentence (3) in the IT-TB. ..... 172
A. 3 'Tree' table for sentence (3) from the IT-TB. ..... 175
A. 4 'Path' table for sentence (3) from the IT-TB. ..... 175
A. 5 'Inpath1', 'target1', 'inpath2', and 'target2' tables for sentence (1) from the IT-TB. ..... 179
A. 6 'Path' table for sentence (1) from the IT-TB. ..... 180
A. 7 Different types of structures in the SC lexicon. ..... 182
A. $8 \quad S C F_{1}$ structure of reservo in sentence (4). ..... 183
A. 9 Frequency of the verb $d o$ with its $S C C_{2_{\text {morph-voice }}}$ structures in the IT- TB lexicon. ..... 185
A. 10 Five most frequent types of SCC and their frequencies for each author in our SC lexicon ..... 186
A. 11 Morphological realisations of verbal arguments in the lexicon. Column 'Sb': the rows headed 'acc' and 'abl' refer to subjects in accusative + infinitive constructions and in ablative absolutes, respectively; 'ind' and 'sbjv' refer to subject clauses whose subjects are expressed as verbs in the indicative and subjunctive mood, respectively. For the other cases, see the text ..... 187
A. 12 Ten most frequent verbs in the LDT and in the IT-TB, and counts of their SCSs in the lexicon. ..... 188
A. 13 Correlation coefficient $r_{s}$, test statistic $S$ (the sum of rank differences) and $p$-value for each correlation test. ..... 191
B. 1 Example corpus frequencies for the subjects of $f l y$. ..... 195
B. 2 Description lengths of the five TCMs for the thesaurus in figure B. $1(\mathrm{Li}$ and Abe, 1998). ..... 198
B. 3 Constructed example of dataset recording the author and the verb lemma for each verbal occurrence observed. ..... 201
B. 4 Frequency table of the dataset represented in table B.3. ..... 201
B. $5 m \times n$ table, with $m$ rows (observations) and $n$ columns (variables). ..... 202
B. 6 Constructed frequencies of the subjects of $f l y$ and $b u z z$. ..... 204
B. 7 Euclidean and cosine distance measures between three pairs of nouns from table B. 6 . ..... 204
B. 8 Cosine distance measures between all pairs of nouns from table B. 6 . ..... 206
B. 9 Clusters obtained at each step from a single linkage clustering and cosine distance on data from table B.6. ..... 206
B. 10 Interquartile Range (IQR) and Median Absolute Deviation (MAD) of the SP probability distributions given by our WN-based system, for each verb in the evaluation set. 'Accido2' and 'dico2' refer to the 'fall' and 'say' senses of the two verbs, respectively. ..... 220
B. 11 Mean, median and third quartile of the SP probability distributions given by our WN-based system, for each verb in the evaluation set. 'Accido2' and 'dico2' refer to the 'fall' and 'say' senses of the two verbs, respectively. 221
B. 12 Frequency and random filler for each verb in the evaluation set, followed by a binary value indicating whether the SP probability of the random filler is higher than the mean, median and third quartile, respectively. 'Accido2' and 'dico2' refer to the 'fall' and 'say' senses of the two verbs, respectively.
B. 13 Comparison between the frequencies and the SP probabilities of the maximum-frequency and minimum-frequency fillers extracted from the parsed corpora. The fourth column contains the maximum frequency of the fillers; the fifth column contains the probability of the maximum- frequency filler; the sixth column contains the mean of the minimum- frequency fillers, the seventh column contains the mean of the prob- abilities of the minimum-frequency fillers; the eighth column contains the binary answers to the question "is the probability of the maximum- frequency filler higher than the mean of the probabilities of the minimum- frequency fillers?"; the ninth column contains the mean of the probabil- ities assigned to the fillers; the tenth column contains the mean of the probabilities assigned to the "non fillers"; the eleventh column to the question "is the mean of the probabilities of the fillers higher than the mean of the probabilities of the "non fillers"?" ..... 224
B. 14 Corpus frequency and SP probabilities for the 'A_(in) $\mathrm{Obj}[\mathrm{acc}]$ '-arguments of intro 'enter'. ..... 225
B. 15 Correlation coefficients $r, p$-values and statistics of the Spearman corre- lation test on the corpus frequency distribution and the SP probability distribution given by our model. ..... 228
B. 16 Correlation coefficients $r, p$-values and statistics of the Spearman corre- lation test on the corpus frequency distribution and the SP probability distribution given by the distributional version of our model. ..... 229
C. 1 Frequency counts of two constructions in Plautus and Thomas in our corpora. ..... 232
C. 2 Expected frequency counts from table C.1 ..... 233
C. 3 Example containing values for car speed and distance to stop. ..... 236
C. 4 Estimated coefficient, standard error, probability of 'construction' and effect size of each predictor in the model. ..... 242
C. 5 Frequency table of a constructed dataset with verbal tokens counted by verb lemma and author. ..... 244
C. 6 Matrix of row profiles from table C.5. ..... 245
C. 7 Matrix of expected frequencies from table C. 5 under the hypothesis of independence. ..... 246
C. 8 Singular values of the MCA for the second investigation on the interaction between 'construction' and 'author'. ..... 251C. 9 MCA Diagnostics of the column points for the second investigation onthe interaction between 'construction' and 'author'. For each profile, thetable contains its mass (second column), the quality of its display in thesubspace (third column), its inertia in permills of the total inertia (fourthcolumn), its coordinates in each dimension (fifth and eight columns), the(relative) contribution of the principal axes to the point's inertia (sixthand ninth columns) and the (absolute) contribution of the point to theinertia of the axes in permills of the principal inertia (seventh and tenthcolumns)255
C. 10 Singular values and inertia explained from the MCA for the second in-vestigation on the interaction between 'construction', 'era', 'preverb','animacy', and 'class'.255
C. 11 MCA Diagnostics of the column points for the second investigation on the interaction between 'construction', 'era', 'preverb', 'animacy', and 'class'. For each profile, the table contains its mass (second column), the quality of its display in the subspace (third column), its inertia in permills of the total inertia (fourth column), its coordinates in each dimension (fifth and eight columns), the (relative) contribution of the principal axes to the point's inertia (sixth and ninth columns) and the (absolute) contribution of the point to the inertia of the axes in permills of the principal inertia (seventh and tenth columns).
C. 12 Singular values of the MCA for the third investigation. ..... 257
C. 13 MCA Diagnostics of the column points for the third investigation. For each profile, the table contains its mass (second column), the quality of its display in the subspace (third column), its inertia in permills of the total inertia (fourth column), its coordinates in each dimension (fifth and eight columns), the (relative) contribution of the principal axes to the point's inertia (sixth and ninth columns) and the (absolute) contribution of the point to the inertia of the axes in permills of the principal inertia (seventh and tenth columns).
C. 14 Frequency counts and expected frequencies of the different combinations of the values of 'era', 'animacy' and 'preposition' in our data. ..... 259

## Chapter 1

## Introduction

This thesis explores an original, computational approach to the study of Latin verbs. We present a subcategorization (SC) lexicon, which is a new corpus-based lexical resource that records morphological, syntactic and semantic information on the argument structure of Latin verbs. The morpho-syntactic and lexical features of the lexicon are associated with the verbal arguments and were extracted from two morpho-syntactically annotated corpora: the Latin Dependency treebank (Bamman and Crane, 2007) and the Index Thomisticus Treebank (Passarotti, 2007b). The semantic information contained in the lexicon regards the verbs' selectional preferences (SPs) - i. e. the lexical-semantic preferences of verbs on their arguments - and was acquired by a computational model that uses both the morpho-syntactic data from the lexicon itself and the lexical-semantic taxonomy of Latin WordNet (LWN; Minozzi 2009). In addition, we present a case study on Latin spatial preverbs where we exploit the data from the lexicon and propose the use of new quantitative methods in Latin linguistics.

The audience we address includes both Latin linguists and computational linguists. Indeed, the SC lexicon can be exploited for synchronic and diachronic studies on various phenomena regarding Latin verbal arguments, as we show in the case study on preverbs. On the other hand, the computational model for SPs, inspired by the work reported in Alishahi (2008), addresses some challenges specific to the Latin data, which makes it interesting for its methodological contribution to computational linguistics as well as for its potential applications.

In this chapter we will briefly sketch the computational approach we propose (section 1.1). Then, we will state our aims (section 1.2) and the scope of our work (section 1.3). Finally, we provide an overview of the rest of the thesis (section 1.4).

### 1.1 Computational linguistics for Latin

The interest in developing computational resources and methods for English and other modern languages is motivated, among other things, by their importance to applications in various areas of society with important commercial uses, such as language technology.

The picture might seem different for ancient languages since the lack of commercial interest has led to a different technological development compared with major contemporary languages. This difference is related both to the peculiarities of the texts analysed by scholars and to the different purposes of linguistic research. In Latin lin-
guistics the connection with the philological tradition and the literary studies has been strengthened by the typical academic nature of the scientific investigation.

However, language technology for ancient languages has gained interest in the research community for Cultural Heritage applications, as testified by the LaTeCH (Language Technology for Cultural Heritage, Social Sciences, and Humanities) workshop organized within several computational linguistics and articifial intelligence conferences in the past years. ${ }^{1}$

Indeed, if we go back to the origins of computational linguistics, we find that it was precisely Latin the language chosen by father Roberto Busa S. J., universally recognised as the pioneer in the use of computers in the humanities because of his Index Thomisticus (IT), the first electronic corpus (see section 1.1.1.1). In spite of this important precedent, over the years computational linguistics for Latin has not been able to keep up with the goals achieved for widely used modern languages like English. The lack of computational tools and electronic corpora morphologically, syntactically and semantically annotated, is thus a challenge for any research aiming at exploring the new field of computational linguistics for Latin.

The present investigation places itself in this area and proposes the use of advanced methods developed in computational linguistics to enrich the material offered by tradition and to reach new goals in linguistic research on Latin.

In this section we aim at defining the scope of the computational approach we propose. First, section 1.1.1 illustrates what a computational approach to linguistics implies and shows how such an approach can contribute to advances in Latin linguistics; second, section 1.1.2 stresses the low-resourced status of Latin and states the need for more computational resources for Latin.

### 1.1.1 Computational approaches

Computational linguistics aims at designing, implementing and applying computational models for natural languages. This includes developing computer systems that automatically process languages (Natural Language Processing; NLP), as well as acquiring, modeling and representing meaning (computational semantics), building, enriching and using corpus data (corpus linguistics), and much more.

In this section we will outline some of the features of a computational approach to Latin, showing how advances in Latin linguistics can be achieved through this strategy. First, we will illustrate the role played by resources like corpora in computational linguistics and show how a computational approach to corpora suits research in Latin linguistics (section 1.1.1.1). Then, we will give a few examples of the potentialities of computational methods in general, which also hold for Latin (section 1.1.1.2).

### 1.1.1.1 Corpora

Linguists can rely on native speakers' intuitions for their studies on extant languages. In the case of dead languages such as Latin, access to linguistic judgements by native speakers is denied: this has justified the recourse to textual evidence throughout the history of Latin linguistics and Latin philology, as well as historical linguistics and philology in general. Traditionally, this evidence is collected either manually by looking directly at the original works, or indirectly by referring to resources like dictionaries, thesauri, lexicons.

[^0]Starting from the middle of the last century, large collections of texts have been collected in more formalised units called corpora. ${ }^{2}$

One of the main characteristics of corpora is that they are typically made available to a wide scientific community to be used for different purposes and studies. Moreover, compared with texts collected for a specific investigation by a scholar or a group of scholars, a corpus is usually larger in size. A few examples of corpora from different periods will illustrate the question of size. The first elecronic corpus was started in the late 1940s by father Roberto Busa S. J. in the form of Thomas Aquinas' opera omnia: the Index Thomisticus (IT; Busa 1980). The IT consists of around 11 million words, a huge figure for those times. The Brown Corpus (one million words) was created in the 1960s at Brown University in the United States, while the British National Corpus ( 100 million words) was built in the 1990s. The largest corpus for Italian, La Repubblica corpus, contains 450 million words extracted from the homonymous Italian newspaper.

As Tognini-Bonelli (2001, pp. 2-4) points out, a text is to be read horizontally, from left to right, within its unique communicative context. On the other hand, a corpus collects several communicative events; rather than being read consecutively, a corpus is typically searched by a keyword (a lemma, a word form, a syntactic feature, etc.); it can thus be conceived of as a list of concordances and be read both vertically, by examining the paradigmatic dimension of the word, and horizontally, by looking at the single local contexts of the word.

For these reasons, corpora offer precious resources for all those linguistic studies which require textual data. In particular, corpora are extremely valuable in Latin linguistics, since they provide (large) amounts of texts in a standard format and are potentially available to the whole scientific community. The availability of such resources makes it possible to carry out corpus-based studies which are replicable, adding a high scientific integrity to them.

Corpora are also the main source of data for most research in computational linguistics. Here we will focus on one specific way to process corpora in computational linguistics: annotation.

Annotation Several digital editions of texts are nowadays available to the scientific community working on Latin. Such raw text corpora, together with concordances, provide important information on the distribution of word forms in a text. However, they do not allow more advanced explorations. For example, they do not allow us to search the contexts of all word forms of a given lemma (such as puellae 'of a/the girl', 'to a/the girl', '(the) girls', puellarum 'of (the) girls', ...for the lemma puella 'girl'), the syntactic roles (such as all subjects of fero 'bring'), or all words semantically related to a given word (such as all synonyms of domus 'house'). This type of rich information is encoded in the text through annotation, which is part of the work done by computational linguists. Annotating a raw text makes explicit different kinds of information that are implicit. A corpus can be annotated at different levels and a branch of research in computational linguistics works with annotation by developing techniques for manual and automatic annotation. See chapter 3 for an overview on annotation tools and resources for Latin.

An example of manual annotation for the Latin sentence sese iactabit audacia 'audacity will throw itself' is given in table 1.1. This table shows a series of attributes for each of the three word forms included in the sentence: "aut" indicates the author of the sentence, "id" identifies the word instance, "form" gives the form of the word,

[^1]aut="Cicero" id="1" form="sese" lem="se" c_m="acc" syn_fun="object" anim="inan"
aut="Cicero" id="2" form="iactabit" lem="iacto" c_m="ind" syn_fun="predicate" sem_cl="throw" aut="Cicero" id="3" form="audacia" lem="audacia" c_m="nom" syn_fun="subject" anim="inan"

Table 1.1: Example of annotation of the Latin sentence sese iactabit audacia "audacity will throw itself".
"lem" is its lemma, "c_m" its case (for nouns) and mood (for verb forms), "syn_fun" its syntactic role in the sentence, "sem_cl" the semantic class of the verb, and "anim" the animacy of the nouns.

If we aim at annotating large corpora, the manual procedure quickly becomes less and less feasible. Automatic annotation can be combined with manual annotation so that the automatic output can be further manually checked and corrected. Systems that automatically annotate corpora are called NLP tools. Both the output of NLP systems and manually annotated corpora can be directly used in various applications because they provide useful data for linguistic explorations of texts. We will refer to this aspect as corpus-based approaches because they employ corpus data in order to answer research questions relative to specific phenomena.

Qualitative and quantitative approaches Corpus-based studies can either aim at a manual inspection and analysis of the corpus data (so-called qualitative approaches) or they can use statistical techniques to deal with these data (so-called quantitative approaches).

The former include the approach traditionally followed in Latin linguistics, where the texts were typically collected, selected and analysed based on a case-by-case philological analysis. However, the availability of large corpora in computational linguistics makes this approach less and less viable for practical reasons.

The size constraints leave room for quantitative approaches. Statistical techniques are designed to account for different corpus sizes and allow us to precisely detect and quantify the role played by the variables encoded in the data, making our results more objective. Chapter 7 gives an example of a corpus-based quantitative study on Latin preverbs and illustrates the characteristics of our quantitative approach.

### 1.1.1.2 The methodological contribution of NLP techniques

Besides providing data ready to be used for corpus-based studies, NLP tools can provide new valuable knowledge on how language works, and confirm theoretical intuitions on language. We will refer to this as the methodological contribution of NLP tools in computational linguistics. Now, we will give three examples of how NLP tools can shed new light on our understanding of natural languages.

Systems for automatic syntactic parsing produce the syntactic analysis of a sentence; this is extremely useful in a number of computational applications. In addition, syntactically analysed corpora offer a great opportunity to corpus linguists interested in investigating various aspects of syntax. However, this is not their only possible use: we will now show that research on parsing can also represent an advance in the theory of syntax.

Parsing can be achieved with so-called rule-based parsers or with statistical parsers. Rule-based parsers use linguistic heuristics, e. g. a noun before a verbal form in English is often the subject of the predicate, apart from identifiable exceptions. On the other hand, statistical parsers are trained on annotated data, i.e. collections of sentences which have
already been syntactically analysed; hence parsers are able to "learn" recurring patterns in the training data and recognise them in new unseen corpora by referring to some selected features, such as Part of Speech (PoS), position in the sentence, case, etc. Our point here is that comparing the performance of a rule-based parser with that of a statistical parser can highlight the most important factors affecting the distribution of syntactic functions in a sentence and tell us which model is closer to the actual syntactic distribution of words.

Another example is offered by research in computational semantics. The so-called "syntax-semantics interface" theoretically assumes that the semantics of verbs partially affect their syntactic behaviour (Levin, 1993). In order to test or exploit this hypothesis, several experiments have been carried out on grouping verbs into classes based on their syntactic properties, for example their arguments: cf. Schulte im Walde (2003) and Lenci et al. (2008), among others. The obtained classes turn out to share certain semantic properties in some cases: the extent to which this is true certainly helps detect which properties are actually relevant for verb semantics and the way these properties vary across languages.

The last example we will give concerns cognitive models of language. These models aim at automatically processing language by trying to mirror the way human brain works. For instance, Alishahi (2008) proposes a model for acquisition of argument structure and SPs in children. We will adapt this model to the case of Latin verbs, thus obtaining a cognitively plausible system for automatically acquiring SPs from corpora.

These three examples show that a computational approach such as ours is not just a tool which offers new and richer data available for the linguistic analysis; it also provides a new way of looking at language. By introducing this new way to investigate Latin, this aspect represents a radical methodological originality of our work.

### 1.1.2 A less-resourced language

We have seen how computational linguistics views corpora and how this can serve the interests of Latin linguists; in addition, we have stressed the methodological potentialities of a computational approach to Latin. Here we want to underline how important it is to build resources for Latin by showing what motivates its less-resourced status.

Latin is a high-resourced language when we consider the number of available texts and traditional lexical resources such as dictionaries. However, if we take a computational point of view, Latin is a comparatively low-resourced language (McGillivray et al., 2009; Passarotti, 2010).

As far as NLP tools for Latin are concerned, experiments on statistical PoS tagging (i.e. assigning the PoS to the word tokens in a corpus) and parsing are being carried out, 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). In addition, a lexical database - Latin WordNet (LWN; Minozzi 2009) - is being compiled. See chapter 3 for a more detailed overview on computational resources and tools for Latin.

Both the number and the size of these resources are relatively small when compared with the corpora and the lexicons available for a modern language like English: cf. for example the sizes of Princeton WordNet (Miller et al. 1990; 155,000 lemmas) and the Penn treebank (Marcus et al. 1993; 2.9 million words). Concerning semantic processing, no semantically annotated Latin corpus is available yet.

For all these reasons, development of NLP for Latin fills a substantial gap in the scientific landscape. Given the time and the energy necessary to manually build such
resources, existing corpora and tools offer great advantages. Moreover, thanks to their language-independent nature, existing automatic lexical acquisition methods and NLP systems should thus be exploited. An example of this attitude is given by our method for automatic acquisition of SPs, which exploits two existing treebanks (in the form of our lexicon) and LWN.

### 1.2 Aims

After the previous introduction on our computational approach and its potential, here we want to summarise our aims. These can be classified into three domains, corresponding to the three content chapters: new methods for (Latin) computational linguistics, new resources available to Latin linguists, and a case study showing a combination of the two.

New methods The methodological aim of this thesis concerns both Latin linguistics and computational linguistics.

In section 1.1.1 we showed how Latin linguistics could benefit from a computational approach. Indeed, our first aim is to propose a methodological innovation in Latin linguistics by presenting a system for acquiring semantic information from corpora which do not explicitly contain semantic annotation. This way we want to propose an alternative analysis to the manual analysis of the data traditionally performed in Latin linguistics studies. This aim is realised in chapter 5 , where we describe our model for SPs of Latin verbs.

However, we believe that our method for semantic processing of Latin can also be of interest to computational linguists. Computational models are usually tested on English in order to compare their performance with previous work. Applying these models to a different language is not a straightforward operation. In the case of an ancient language like Latin, the challenge is made harder by the lack of resources and the peculiarities of Latin morpho-syntax, as well as the special philological and literary interest surrounding the data. As a consequence, new strategies need to be devised in order to adapt the existing systems and to achieve good results with little resources. These strategies can be employed for other languages facing the same kind of problems and offer a methodological term of comparison for the computational linguistics community.

New resources In section 1.1.2 we stressed the less-resourced nature of Latin. In this thesis we present a new resource which partially fills this gap thus addressing our second aim: the SC lexicon described in chapter 4, enriched with SPs acquired as described in chapter 5.

A case study Finally, the corpus-based diachronic analysis of Latin spatial preverbs presented in chapter 7 is intended as a case study to show one of the possible uses of our SC lexicon for analysing a linguistic phenomenon. In addition, the originality of our corpus-based approach is methodological and lies in its quantitative nature.

Therefore, our case study combines the use of a new resource for Latin with our third aim, which consists in a methodological contribution to Latin diachronic linguistics towards quantitative approaches.

### 1.3 Scope. Corpora and language

In the present work we refer to two Latin treebanks, containing works by a number of different authors, from the Classical era to the Medieval era. This corpus is the source of linguistic data for building our SC lexicon and acquiring the SPs included in the lexicon, as well as for carrying out the corpus study on preverbs.

This section defines the scope of our research by discussing our point of view on the relation between corpora and language and the role played by our corpora in the work we propose. Section 1.3.0.1 introduces the corpora and discusses their role in the case study, while section 1.3.0.2 touches on how we used them for the SC lexicon and the SPs.

### 1.3.0.1 Corpora for the case study

A corpus is defined in McEnery and Wilson (2001, p. 32) as "a finite-sized body of machine-readable text, sampled in order to be maximally representative of the language variety under consideration". Unlike a single text, a corpus is naturally oriented towards generalizing the information contained therein to the language as a whole.

Corpora are the foundation of any corpus-based study, such as our case study on Latin preverbs. Corpus-based studies, unlike text-based studies, aim at describing a larger portion of language than that included in the corpora, as we saw in section 1.1.1.1. In statistical terms, this larger portion of language is called a population, while a sample is the subset of the population providing the data evidence for the analysis.

This section answers the question of how good our corpora are for our case study, i. e. how well they reflect the composition of the population. We will face some of the issues related to sampling and representativeness of corpora in our research by summarising the properties of "good corpora", presenting the composition of our corpora, outlining the difficulties of defining the Latin language in general, and finally suggesting an operational definition for it which is suitable in our case.

Issues on representativeness Two criteria to evaluate how suitable a corpus is for a linguistic analysis are its size and its content.

First, if we consider a corpus as a sample, its size helps decide if it is a good corpus: in fact, the number of observations and types that are in the population but are not captured by the sampling procedure (unseen events) tends to decrease as the size of the sample grows. In other words, a larger corpus would allow a better modeling of the variability in the sublanguage of the study: cf., for example, Gorard (2003).

Second, the way the corpus was built may affect the possibility of a good generalization to the sublanguage studied. A still popular criterion for a good selection of texts in a corpus is the notion of representativess: a representative corpus ideally accounts for the whole range of variability in the features of a language: for an introduction to the problem, cf. Lenci et al. (2005). The degree to which a corpus is representative of the general variety of a language has been correlated with its being balanced from the point of view of the chronological period the texts belong into, their genre, modality, domain, etc.

Further methodological remarks In spite of the efforts to build balanced and representative corpora, it has been acknowledged that representativeness is an ideal limit and all corpora show a certain degree of arbitrariness (Lenci et al., 2005, p. 41).

| corpus | author | time | age | genre | word tokens |
| ---: | :--- | :--- | :--- | ---: | ---: |
| Plautus | Plautus | ca. $254-184 \mathrm{BC}$ | early republic | comedy | 17011 |
|  | Cicero | $106 \mathrm{BC}-43 \mathrm{BC}$ | late republic | rhetoric | 6229 |
|  | Caesar | $100 \mathrm{BC}-44 \mathrm{BC}$ | late republic | historiography | 1488 |
|  | Sallust | $86-34 \mathrm{BC}$ | late republic | historiography | 12311 |
| LDT | Vergil | Propertius | $70 \mathrm{BC}-19 \mathrm{BC}$ | Augustan | epic |

Table 1.2: Composition of the reference corpora for our study by author: dates of birth and death of the author, literary age, genre and number of word tokens are provided.

If the corpora are meant as a true sample from a sublanguage, any model of these data will generalize over the whole population through estimates, and will consequently contain some uncertainty. In this perspective, any choice of the linguistic sample is bound to be improvable, thus no choice is perfect. We can then turn the focus from an absolute value of representativeness to a measure quantifiable in statistical terms and hence translate the question of the goodness of the sample into the following:

How much uncertainty are we able to remove from the data at hand?
If we view a statistical and data exploration analysis as a procedure which tries to reduce the degree of uncertainty present in the data, then the previous question can be answered through some statistical modeling such as regression analysis, as shown in chapter 7.

A preliminary methodological conclusion could be phrased as follows. In addition to choosing the corpora in such a way that they are as representative as possible of the sublanguage we are studying, we can choose the statistical techniques in such a way that uncertainty is maximally reduced in our data and it is as safe as possible to generalize the results to the whole population. Now we are going to investigate how these remarks can be applied to our corpora. First, we will show the composition of the corpora and discuss the relation between their degree of representativeness and the definitions of Latin as a language. Then, we will present our view on the relation between our corpus and the language it aims at modeling.

Composition of our corpora Table 1.2 records the composition of the corpora we used in our study in terms of the number of tokens, alongside chronological and genre details. From this table we can classify our data into six literary ages and eight genres spread over 10 authors, although this is only one of all possible ways to view the data. The chronological gaps between Petronius, Jerome and Thomas are substantial with respect to the relative homogeneity observed for the other authors. Moreover, the number of tokens is considerably higher in Thomas than in all the other authors.

These observations allow us to conclude that ou darta are not a balanced sample by genre and era. However, as we anticipated in the previous paragraph, we have access to statistical techniques that are able to accurately assess the impact of this imbalance. In addition, we have some methodological and pragmatic reasons to believe that the choice of the corpora can still be considered to be good, given the practical circumstances of our investigation, as we will see now.

Which Latin? The main requirement for a good corpus of Latin is its representativeness, as we pointed out in section 1.3.0.1. The next question concerns what is a corpus of Latin supposed to be representative of. Generally speaking, we expect a corpus-based study of Latin to aim at offering a (possibly new) view on a linguistic phenomenon regarding the Latin language in general, or at least a clearly limited portion of it. In both cases, it is necessary to have an idea of what we mean by the Latin language.

The difficulty of defining the Latin language has been acknowledged by a number of scholars. Among these, Poccetti stresses how any account of the history of the Latin language must deal with its manifold nature and a series of dichotomies: local vs. universal, dead vs. extant, spoken vs. written, vulgar vs. literary (Poccetti et al., 1999, p. 9-30). Latin can in fact be identified as the language of a particular city, while at the same time representing a universal heritage, extending its time span long after the end of the Roman empire. This double identity is related to a definition of the chronology for Latin, which is an issue in itself: the Indo-European origins of Latin have been widely accepted, whereas a number of different eras have been suggested as marking its end as a spoken language, ranging from the $3^{\text {rd }}$ century A. D. to the $8^{\text {th }}$ century (Poccetti et al., 1999, p. 19), depending on whether we focus on the decline of the Roman empire's unity, the emergence of national languages, or the different political and cultural changes which affected the history of Europe.

In turn, the chronological matter brings with it the question of Latin as a dead language. Such a characterization originated in the area of Renaissance Italian philology and has been emphasized by a long scholastic tradition, which still represents the first filter through which most people approach the Latin language. This tradition has favoured a normative view on Latin grammar, restricted to a limited time span - the classical era and its literary sources, Cicero in primis - considered the qualitative peak defining the identity of Latin (Giacomelli, 1993, p. 13). On the other hand, the vitality of Latin has expressed itself in its role as a language for international communication throughout the centuries, as well as in its continuations in the Romance languages and in the technical domain of a number of languages not directly derived from Latin, such as English. In addition, Latin texts are still produced, for example, by the Vatican, by certain linguistic circles working on Latin and by the Latin version of the online encyclopaedia Wikipedia.

Finally, the contrast between written Latin trasmitted by the available sources and spoken Latin reconstructed from indirect evidence is partially connected with the fuzzy dichotomy between literary Latin and vulgar Latin, which has led to a number of studies and debates, particularly for the relevance of non-classical Latin in investigating the origin of Romance languages: cf. e. g. Rosén (1999, pp. 15-18).

All these dichotomies are mirrored in the texts that are available for Latin. According to a classification of dead languages based on their sources (Mayrhofer, 1980; Untermann, 1983), Latin can be considered a large-corpus language (Großcorpus-Sprache), because of the amount of its available texts. In spite of these factors, as pointed out, for example, in Poccetti et al. (1999, p. 28-29) and in Rosén (1999, p. 31), the corpus of texts available for Latin is extremely heterogeneous, displaying an uneven distribution along various dimensions, including the diachronic one. For example, the sources for archaic Latin - consisting of a few epigraphs and literary fragments from the $6^{\text {th }}$ century B. C. - are very sparse and contrast with the increasing number of texts available from the $3^{\text {rd }}$ century B. C. on.

The overall body of Latin texts, hence, contributes to giving us a complex picture of the Latin language. The linguistic tradition nonetheless recognises how this picture is not complete, not only because of the fragmentary character of the texts or the lack
of evidence from certain genres, registers and eras, but also because it obviously leaves out all direct accounts of spoken Latin.

Spoken Latin and vulgar Latin According to Bloomfield, "writing is not language, but merely a way of recording language by means of visible marks" (Bloomfield, 1933, p. 21). Therefore, spoken evidence would be our best evidence for all languages, including Latin. However, as we will explain, this is particularly difficult in the case of Latin, given the nature of our sources.

The word "vulgar" is often used to indicate the spoken component of Latin, although the definition of vulgar Latin is controversial: see, for example, Väänänen (1971). Beyond the different views on its characterization, we can state that "vulgar Latin" generally refers to the collection of linguistic features not associated with standard Latin, i. e. the classical and literary norm and the high register.

Our knowledge of spoken or vulgar Latin is anyway mediated by the written texts we are provided with. Along with inscriptions and papyri, a list of authors have been proposed as sources for vulgar Latin: Plautus' plays, Cicero's letters, Cena Trimalchionis by Petronius. Nevertheless, scholars tend to agree that these sources, alongside a literary genre, display different aspects of the vulgar nature of the language, ranging from domain specificity to social class, background and target audience. We should then consider these texts as showing some vulgar elements, rather than calling them vulgar tout court (Poccetti et al. 1999, pp. 21-27, Rosén 1999, p. 19).

Spoken Latin is thus largely unknown and unknowable. However, we do not see reasons to dismiss the value of corpus studies as source of insight to a language merely because they rely on written sources. We believe that no matter how high the value of spoken language is with respect to written language, texts are still the best evidence we can have for modeling Latin.

Thomas The remarks just outlined also explain our choice to include Thomas as part of our corpus.

Some might object that Thomas was not a native speaker of Latin. From what we said in the previous paragraph, we believe that relying on traces of spoken Latin as the only basis for a linguistic analysis would lead us to exclude important contributions. This is in line with Croft (2000)'s position, according to which a language is "the population of utterances in a speech community"; an utterance is defined as a token, i. e. "a particular, actual occurrence of the product of human behavior in communicative interaction (i.e. a string of sounds), as it is pronunced, grammatically structured, and semantically and pragmatically interpreted in its context" (Croft, 2000, p. 26). In other words, this approach moves the focus from who writes (the individual authors) to what is written (the texts) and considers all utterances as equally relevant for the study of language.

In addition, given the diachronic aim of our case study, we think that the data for Thomas are essential for detecting a trend in time. Of course, while interpreting the results, we will have to take into account the differences between the various authors analysed.

Finally, there is also a statistical-methodological motivation that led us to this choice. The corpus for Thomas is large compared with the other authors, and we use statistical techniques that account for these size differences. On the other hand, if we had excluded Thomas, we would have had a much smaller corpus, which would have led to more uncertain results and a larger margin of error in our estimations.

An operational definition of language We just outlined the problems arising when defining the time span of the Latin language, the extent to which it can be considered a dead language, and the degree to which we have access to its spoken variety.

Given all these difficulties, the best option for defining Latin seems to be to overcome any dichotomy and consider Latin as the outcome of a series of particular historical, geographical and cultural circumstances, necessarily leading to an inhomogeneous linguistic system where elements from different areas and registers met and were only partially transmitted by the sources.

Here we propose a way of defining the language variety of Latin which we want to investigate; this definition justifies our use of the corpus evidence for the case study. The corpora we used are a subset of a larger unannotated corpus; this larger corpus could be thought of as the whole Perseus Digital Library joined with the Index Thomisticus and all the plays by Plautus. We consider our sublanguage to be this larger corpus.

An advantage of our choice resides in its possibilities for future research, which introduce an important methodological innovation of our study. In fact, the statistical analyses we carried out are able to make predictions on the population by using the sample at our disposal. We can hypothesize that the corpus constituting our sublanguage will be annotated in the future, thus providing a basis to test the validity of our predictions. The syntactic annotation can for now be approximated by an automatic parsing, which is in fact in progress for the Index Thomisticus and the Perseus Digital Library (Bamman and Crane, 2008; Passarotti and Ruffolo, 2009). Hence, in the future it will be possible to evaluate our study by comparing its results with the parsed data.

Randomness So far, we have stressed the importance that a corpus is a good sample for the sublanguage we aim at modeling by focussing on the relation between the corpora and our definition of a sublanguage of Latin. Now, we want to define the conditions under which our corpora are a good sample for the sublanguage of Latin we just defined from the point of view of statistical modeling.

The best way to obtain a sample from a population is by random sampling (Gorard, 2003; Rasinger, 2008), where each member of the population has the same chance of being included in the sample. Evert (2006) introduces a new way of looking at the problem of random sampling in linguistics. According to Evert, language is not inherently random, in the sense that language utterances are not random bags of words. If linguistic samples are not random at the word level, the randomness may be seen at a different level. To explain this, Evert (2006, p. 180) thinks of language as a large library:

The selection of a particular corpus as the basis for a study among all the other language fragments that could also have been used is like picking an arbitrary book from one of the shelves in the library. It is this choice which introduces an element of randomness into corpus frequency data; and it is this element of randomness, in turn, that needs to be accounted for by the methods of statistical inference.

Linguistic data do not consist of words that are put together randomly; rather, they are combined based on syntactic rules, conventions, pragmatic considerations, etc. This causes serious problems for statistical tests that assume randomness. However, if we consider the corpus compilation level, we might be justified in assuming an element of randomness while using these tests, provided that we create the hypothesis and the population with this new perspective in mind. In other words, the conditions under
which we can regard our corpora as good samples translate into the following question: is the choice of the texts in the corpora biased somehow? If not, then the corpora can be considered a random sample at the text level, in the sense specified by Evert. See Jenset (2010, pp. 43-45) for a discussion of these issues in historical linguistics.

We will now show that random sampling holds for the purposes of our study because the corpora were not chosen with the aim of a particular linguistic study in mind. Indeed, the IT-TB was built by collecting the concordances of the lemma forma 'form' in the Index Thomisticus; this means that the choice of the IT-TB is not biased by the purpose of studying Latin preverbs, since we can assume that the occurrences of forma are randomly distributed with respect to the occurrences of PVs. For what concerns the LDT, although we might recognise that well-known authors are included therein, no particular selection criterion appears in the documentation. Finally, the two plays by Plautus were chosen for this particular study. However, there is no study proving that these plays contain some kind of biased evidence for Latin PVs. These considerations allow us to consider our corpora as a good sample from the population we have defined, provided that we use appropriate statistical techniques that account for the peculiarities of the texts.

### 1.3.0.2 Corpora for the SC lexicon and the SPs

The SC lexicon and the SPs were obtained from the LDT and the IT-TB. Since this part of our work focusses on building a resource and proposing an acquisition system rather than analysing a specific linguistic phenomenon, the corpus data coincide with the sublanguage we want to model. Hence, no generalization from a sample is involved; instead, we aim at the most faithful representation of the reference corpora used.

In the case of the SC lexicon, the features of the corpus are mirrored in the metalinguistic information contained therein, such as author, work, etc. This means that those interested in diachronic analyses of the SC behaviour of Latin verbs can search the lexicon by these variables.

In the SP acquisition phase, we chose to merge together the verb instances from the two treebanks. This was motivated by sample issues, since the small size of the data implies a lot of unseen instances, and this effect would increase if the data from the two treebanks had been kept disjoint. For example, the verb accedo 'access' appears 13 times in the corpus with the preposition ad 'to' followed by the accusative case; this is the distribution of its fillers by author: sportella 'little basket' (1), auris 'ear' (1), cibus 'food' (1), caput 'head' (1) in Petronius, urbs 'city' (2) in Sallust, deus 'God' (1), similitudo 'similarity' (2), nobilitas 'nobility' (1), iuramentum 'oath' (1), ieiunus 'hungry' (1), elementum 'element' (1) in Thomas. All these nouns contributed to the acquisition of SPs for this verb. However, the procedure can be replicated on any selection of authors within the corpora, if diachronically-characterised data are needed: cf. chapter 5 .

### 1.4 Overview of the thesis

After this introductory chapter, the thesis contains the first background chapter, chapter 2, which provides a general overview on the linguistic phenomenon we are investigating: the argument structure of Latin verbs. We review some theories on argument structure in general and go through the peculiarities of the argument structure of Latin verbs. The second background chapter, chapter 3, describes the state of the
art on computational resources and tools for Latin.
The subsequent three chapters are content chapters and introduce the core of our original contribution.

Chapter 4 presents the SC lexicon we extracted from the LDT and the IT-TB, after reviewing previous work on SC lexicons. Then, we describe the annotation of the treebanks, our extraction queries, and how we represented argument structure information in the lexicon. Finally, we provide some statistics on the lexicon by quantitatively comparing the data from the LDT and the IT-TB.

Chapter 5 illustrates our system for acquiring SPs from the two treebanks by using the SC lexicon and LWN. We first review previous work, then describe our model, and finally present evaluation results.

Chapter 7 illustrates our corpus-based quantitative study on Latin spatial preverbs by using the SC lexicon; it is preceded by chapter 6 , which gives an overview on previous research on spatial preverbs and Latin preverbs.

Finally, chapter 8 provides some conclusions, summarises the contribution of our work and suggests future directions of research.

Appendix A contains the technical details of the queries used to build the lexicon.
Appendix B covers some technical aspects useful to understand the details of previous SP systems, presented in chapter 5, as well as ours. In addition, it gives a brief introduction to general concepts such as vector spaces and clustering, assuming no background knowledge from the reader.

Appendix C illustrates the techniques used for the quantitative analysis contained in chapter 7 by outlining the main ideas behind them, again without assuming that the reader knows the topic in advance.

Given the heterogeneous audience we address, which is composed of computational linguists and Latin linguists, we have tried to provide an introduction to the topics that are specific to each discipline. These include the properties of arguments of Latin verbs and their treatment in Latin dictionaries (section 2.3), as well as clustering (section B.5) and linear regression modeling (section C.3). All this may result in an eccess of explanation of very basic concepts for some readers, but hopefully such a feeling will be equally shared by the two groups.

## Chapter 2

## Background

### 2.1 Introduction

As we said in chapter 1, this thesis proposes new computational resources and methods for the analysis of Latin verbs and their arguments. In accordance with this plan, we tackle the main directions of our research in three dedicated chapters: chapter 4 describes the subcategorization (SC) lexicon we built, chapter 5 illustrates our method for acquiring selectional preferences (SPs) for Latin verbs and chapter 7 contains our quantitative analysis on Latin spatial preverbs. Each of these chapters is provided with a literature review: section 4.3 in chapter 4 , section 5.2 in chapter 5 and chapter 6 for chapter 7 .

The present chapter serves as a general background to our object of study, i.e. predicate arguments in general (section 2.2) and, in particular, the properties of Latin verbs in terms of SC and SPs (section 2.3).

### 2.2 Arguments and adjuncts

In this section we define what we mean by 'argument' and 'adjunct'. In section 2.2.1 we discuss and exemplify these notions based on English; this will highlight the typological nature of the problem as well as serving as a contrast with the discussion on Latin provided in section 2.3. Section 2.2.2 reviews some theoretical approaches to verbal arguments and adjuncts.

### 2.2.1 Definitions

Predicates tend to occur with other constituents. Once we specify a predicate whether it is a verbal form, or an adjective or noun with a copula - its co-occurrence with certain constituents can be predicted to some extent: we will refer to these constituents as arguments.

For example, if we consider the English verb give, we expect it to occur with an NP indicating the person giving something and/or a noun phrase for the thing given, and/or an NP or a to-PP for the person receiving the thing. The form of these constituents is determined by the predicate so that, for example, the third constituent we mentioned cannot be expressed as a PP with the preposition in. Give can also occur with other
constituents, such as temporal or spatial expressions, but we cannot as easily predict these co-occurrences based on the verb only.
(1) She gave him the book yesterday/in the garden/because he arrived.

In (1) we have three arguments (she, him and the book) whose presence and form can be predicted based on the verb give; in addition, we have a set of temporal, spatial and causal expressions (yesterday, in the garden, because he arrived), which cannot be predicted based on the verb give only. The presence of the latter constituents is governed by pragmatic or communicative factors, not by the verb; likewise, their form is not determined by the verb: we will call them adjuncts.

The distinction between arguments and adjuncts is a gradience: on the one end of the scale we find obligatory arguments, and on the opposite end we have adjuncts, while the elements appearing in the middle of the scale are less easily distinguishable. Obligatory arguments cannot be missing while keeping the sentence grammatical and the sense of the predicate constant: for example, if the adjective dangerous is taken away from sentence (2), the meaning of the verb consider changes from 'have a certain opinion of something/someone' to 'take into consideration something/someone' (sentence (3)).

I consider you dangerous.
I consider you.
Arguments are optional when they are required by the predicate and their form is determined by it, but they can still be left out: their absence does not affect the grammaticality of the sentence nor the sense of the verb, although it may change its communicative value. For example, the English verb eat has an obligatory subject and an optional direct object, as shown by the grammatical sentences (4) and (5): in these sentences the sense of the verb does not change, altough its telicity does (from telic to atelic). On the contrary, a synonym of eat like devour always requires a direct object.
(4) She is eating an apple.

She is eating.
Given the gradience of the distinction between optional and non-optional arguments, some cases are uncertain. Typical examples concern spatial complements occurring with verbs involving a motion event like come: in the sentence She came home from work at 8 , the complement home could be considered optional or obligatory depending on whether we view it as an integral part of the semantics of come or not.

Several criteria have been proposed for distinguishing optional arguments from adjuncts, but they are not completely satisfactory: see Herbst et al. (1980) and Helbig (1992) for German and Buysschaert (1982); Somers (1984) for English. For example, adjuncts tend to occur as adverbs (I ate there), but this behaviour does not exclude arguments (We went there). Moreover, especially in languages with a relatively fixed word order such as English, changing the position of the arguments affects the focus of the sentence (I can't watch this movie vs. This movie I can't watch), while changing the position of adjuncts usually does not (I ate an apple yesterday vs. Yesterday I ate an apple). The argument/adjunct distinction is in fact a central problem in syntax, whose solution is crucial not only in theoretical linguistics but also in NLP applications, as proven by the different methods proposed for extracting verbal arguments from texts (see section 2.2.2.2).

The fuzzy boundary between arguments and adjuncts is also relevant for a semantic analysis. The arguments of a verbal predicate correspond to the participants of the event described by the verb: these participants are characterized by thematic roles (also called semantic roles or theta roles: cf. Fillmore 1968 and Gruber 1976) based on the role they play with respect to the event: agent, patient, experiencer, theme, source/goal/location, recipient/beneficiary/maleficiary, instrument, etc.

The connection between the morpho-syntactic expression of arguments and their semantic roles is called linking. For example, give in sentence (6) describes an event with three participants: agent (he, syntactically expressed as a subject), recipient (her, first object) and theme (the book, second object).
(6) He gave her the book.

Linking morpho-syntactic expressions of arguments and their semantic roles is not always straighforward, especially for NLP applications, because the same verb can express the same proposition (with different pragmatic and stylistic properties) through different argument configurations, called diathesis alternations. An example is the active/passive alternation (sentences (7) and (8)), where the syntactic function of subject may express the agent in the active sentence, as in (7), and the theme in the passive sentence, as in (8).
(7) The boy ate the apple.
(8) The apple was eaten by the boy.

Several theories of semantic roles exist which use the semantic role labels in different ways: a discussion of this is provided in Levin and Rappaport Hovav (2005, pp. 3549). In order to overcome some of the problems related to the "unclear boundaries" of semantic roles, Dowty (1991) proposes a higher level of generalization and distinguishes them into two sets of so-called proto-roles: proto-agent and proto-patient. Proto-roles are defined in terms of certain semantic properties, such as volitional involvement for proto-agent or undergoing of a change of state for proto-patient. There is a range of degrees to which arguments can be proto-agents or proto-patients: for example, the subjects of kill are more prototypically agents than the subjects of feel are.

Dowty's proto-roles might be cognitively plausible, given the high ability of human mind to disambiguate in context. However, a more fine-grained distinction of roles could be more effective for computational applications. An example of the most fine-grained classification possible is represented by verb-specific descriptions of semantic roles; this classification is used in some lexical resources for English like PropBank (section 4.3.2). For example, the first argument (agent) of the verb decontaminate can be identified as 'cleaner', and the second argument as 'dirty thing, now clean'.

Selectional preferences Another way to look at the semantic interactions between a predicate and its arguments is to consider the arguments' lexical-semantic features. Arguments are selected by verbs according to specific properties, called selectional preferences (SPs; Wilks 1986) or selectional restrictions (Katz and Fodor, 1964).

Unlike semantic roles, SPs are lexical-semantic properties of the arguments, or rather, they are semantic properties of the lexical fillers occurring with a verb in a specific argument role. For example, in sentences (9) and (10) the subject position of fly is filled by plane and bird; it refers to the semantic role of agent and expresses the entity flying. Only certain categories of concepts can play this role (for example birds,
insects, or certain vehicles), and fly restricts the constituents appearing as its first argument to these categories (or prefers constituents belonging to these categories). Consequently, a sentence like (11) would be ungrammatical, or acceptable only in a metaphorical context.
(9) The plane is flying over us.
(10) The little bird flew for the first time.
*The train flew to Rome.

The SPs of a verbal lexeme may vary depending on its senses. For example, the English verb serve has a cooking sense and a commercial one: in the first case it prefers food in its theme position (they served pizza), in the second it prefers geographical entities (this company serves Sardinia). For these reasons, SPs have been used in Word Sense Disambiguation systems in computational linguistics: cf. e. g. McCarthy (2001).

In addition, the SPs of a verb may vary in specificity so that a very broad class may be appropriate for some verbs, while others may require a narrower class. For example the theme position of think is filled by a very wide variety of entities, while the verb diagonalize refers to specific mathematical objects (matrices). In section 5.2 we discuss the problems involved in this generalization step from lexical fillers to broader classes and we give an overview of computational models for acquiring SPs from texts.

From the discussion so far we can conclude that the number and type of arguments a predicate can occur with can be seen as a lexical feature of the predicate and described from a syntactic or a semantic point of view, or with a combination of both. This feature has received different names in different linguistic theories: for example, in English linguistics it has been investigated as part of the predicate's argument structure (Levin, 1993; Levin and Rappaport Hovav, 2005; Goldberg, 2006). On the other hand, in Dependency Grammar (DG) it has been investigated in terms of valency (Tesnière, 1959; Mel'čuk, 1988; Pinkster, 1990).

A term focussing on the syntactic functions of the arguments is subcategorization (SC), which is widely used in Computational Linguistics and NLP research: cf. McCarthy (2001); Korhonen (2002), for example. This name refers to classifying verbs sharing a particular set of semantic arguments into subcategories based on the different syntactic realisation of these semantic arguments. For example, within the class of Latin verbs with semantic arguments agent and location, there is a subcategory whose verbs express these arguments with a subject and an object in the accusative case, such as adeo 'go to, reach' in (12), and another subcategory that uses a subject and prepositional phrase (PP) with ad 'to', such as eo 'go' in (13).
(12)

$$
\text { adit } \quad \ldots \text { ortus } \quad(\text { Ov., Met., I, 779) }
$$

go to:IND.PRS.3SG . . . sunrise:ACC.M.PL
'he rushed to the orient'
(13) ibat res ad summam nauseam (Petron., Sat, 78)
go:IND.IMPF.3SG thing:NOM.F.SG to highest:ACC.F.SG nausea:ACC.F.SG
'The thing was becoming extremely sickening'

These two subcategories would lead to two different patterns: \{Subject, Object[acc]\} and $\{$ Subject, $(a d)$ Object $[\operatorname{acc}]\}$, respectively. In our SC lexicon we will see how we decided to represent such patterns.

### 2.2.2 Approaches

One of the aims of our research is to provide a lexical resource for describing Latin verbal arguments; for this reason, we take a lexicographic approach and focus on those theoretical frameworks which served as background to the valency and SC lexicons which we will review in section 4.3: for each of these theories here we will outline their views on predicate arguments. We will start from DG and its syntactic and semantic view on valency (section 2.2.2.1), since the corpora we use were annotated according to this framework; then, section 2.2.2.2 contrasts DG and Phrase Structure Grammar (PSG) for what concerns the definition of verbal arguments. Finally, section 2.2.2.3 deals with different views on argument structure from a semantic viewpoint.

### 2.2.2.1 DG and Valency Theory

Here we present Valency Theory from a DG perspective, although valency phenomena have also received attention in other frameworks such as Functional Grammar, Lexical Functional Grammar and others. Some valency lexicons referring to the theoretical framework of valency theory will be described in section 4.3.1.1 as part of the background to our SC lexicon.

DG is a comprehensive class of grammatical theories and formalisms tracing their origins to Tesnière (1959). Besides Tesnière's theory of structural syntax, other types of DG exist: Dependency Unification Grammar (Hellwig, 1986), Functional Generative Description (Sgall et al., 1986), Meaning Text Theory (Mel'čuk, 1988), Word Grammar (Hudson, 1990), etc.

A DG-based annotation will be further described in section 2.2.2.1 as the framework for the annotation of the Latin treebanks we use. Here we will go into how DG defines arguments and adjuncts.

The characterization of a predicate in terms of the number and type of its arguments is defined as valency in DG. In order to explain what valency is, it is necessary to outline what dependency is.

DG represents the syntactic structure of a sentence by the dependency relations between its words, where 'word' refers to a wordform, i.e. a lexeme in a specific inflectional form, e.g. speaks or spoke for the lexeme speak (Mel'čuk, 2003). In Tesnière (1959, pages 11-13) we find the following definition of dependency (emphasis in the original).
1.2. - La phrase est un ensemble organisé dont les éléments constituants sont les mots. 1.3. - Tout mot qui fait partie d'une phrase cesse par luimême d'être isolé comme dans le dictionnaire. Entre lui et ses voisins, l'esprit aperçoit des connexions, dont l'ensemble forme la charpente de la phrase. [...] 2.1. - Les connexions structurales établissent entre les mots des rapports de dépendance. Chaque connexion unit en principe un terme supérieur à un terme inférieur. 2.2 - Le terme supérieur reçoit le nom de régissant. Le terme inférieur reçoit le nom de subordonné. Ainsi dans la phrase Alfred parle [...], parle est le régissant et Alfred le subordonné. ${ }^{1}$

[^2]In other words, the dependency relations between the words of a sentence are binary relations between syntactically subordinate words (dependents, also called subordinates or satellites) and the words they depend on (heads, also called governors, regents, or rulers).

Figure 2.1 illustrates a dependency diagram for the English sentence (14).
(14) She ate the apple.


Figure 2.1: A dependency representation of sentence (14).

In this representation the arrows point from the dependent to the head and are labeled with the respective dependency relations: 'Determiner', 'Subject', and 'Object'.

Several criteria have been proposed for defining head-dependent relations in DG, depending on whether we view them as notions at the morphological, syntactic, or semantic level (Zwicky 1985; Mel'čuk 1988; Hudson 1990, 1993).

Valency Although dependents are not defined in DG with respect to verbs only, verbal predicates have a special role in DG, since it is the verb's dependents that determine the presence of the other constituents in the sentence.

Tesnière (1959, p. 102) uses a metaphor borrowed from drama to classify verbal dependents into 'actants' (arguments) and 'circonstants' (adjuncts):

Le noeud verbal [...] exprime tout un petit drame. Comme un drame en effet, il comporte obligatoirement un procès, et le plus souvent des acteurs et des circonstances. Transposés du plan de la réalité dramatique sur celui de la syntaxe structurale, le procès, les acteurs et les circonstances deviennent respectivement le verbe, les actants et les circonstants. ${ }^{2}$

The verb's arguments (actants) form its valency, while adjuncts (circonstants) fall outside the verb's valency. For example, the valency of the verb change requires a subject in sentences like The city has changed since last year. Therefore, this DG-oriented definition of valency focusses on the verb's dependents rather than on the constituents co-occurring with the verb.

From a syntactic point of view valency is thus the number and formal character of arguments a predicate takes. Consequently, a sentence like (15) is considered as showing

[^3]a mono-argumental instance of the verb meet, and a sentence like (16) as showing a bi-argumental instance.

## They have never met.

She met him.
However, we could define valency from a semantic point of view as the number and character of arguments a predicate takes, independently of their formal realisation. For example, we could say that meet requires two "semantic" arguments (the person who meets and the person who is met) and these arguments are syntactically expressed by one syntactic constituent in (15) and by two syntactic constituents in (16). Therefore, the semantic valency can be numerically higher than the syntactic valency since not all semantic arguments have to be syntactically expressed to make the sentence grammatical. Another example is the verb hit, which semantically requires the instrument with which the hitting is done, but the instrument does not have to be syntactically expressed.

Several scholars have attempted to investigate the relation between a syntactic view on valency and a semantic one (Helbig, 1992; Mel'čuk, 2004a,b): we refer to GötzVotteler (2007) for an overview of these theories. The next paragraph illustrates how a specific DG theory deals with the argument/adjunct distinction and valency from a combined morpho-syntactic and semantic point of view.

FGD Functional Generative Description (FGD; Sgall et al. 1986) is a DG framework. It was first theorized by Petr Sgall in the 1960s and further developed by Eva Hajicová and Jarmila Panevová; in the last decade it proceeded towards the creation of computational lexical resources.

In FGD arguments are called 'inner participants' and adjuncts are called 'free adverbials' or 'free modifications': by definition free adverbials, unlike inner participants, can combine with all verbs. Inner participants can be obligatory or optional, depending on whether their omission causes the sentence to be ungrammatical or still grammatical.

FGD is the theoretical framework of the Prague Dependency Treebank (PDT; Böhmová et al. 2003), as becomes evident when considering its annotation scheme. The PDT consists of a part of the Czech National Corpus annotated at three 'layers', all linked together: morphological ( 115,844 sentences), analytical ( 87,913 sentences) and tectogrammatical ( 49,431 sentences). The morphological annotation records the lemmas of the word tokens, as well as their part of speech and other morphological categories. The analytical level represents the dependency tree of each sentence: the nodes of the tree correspond to all and only the words of the sentence, labeled with their analytical function, that is their syntactic relation with their head. Finally, the tectogrammatical level reflects the semantic structure of the sentences. It contains the so-called "auto-semantic" words only, thus excluding punctuation marks and auxiliary elements; these words correspond to nodes in the semantic tree, annotated with their semantic dependency with respect to their governor. This dependency is expressed in terms of specific tectogrammatical functions and is called functor. The main functors are: ACT(or), PAT(ient), ADDR(esse), ORIG(in) and EFF(ect)

In FGD valency is considered to act at the tectogrammatical level and is defined as the ability of certain lexical heads to govern specific functors. This is exemplified by the two valency lexicons based on FGD (PDT-Vallex and Vallex), which are illustrated in section 4.3.1.2. Since the semantic dependency between the elements is associated with its representations at the syntactic and morphologial levels, valency too is related
with the preceding levels of annotation. The first two functors (actor and patient) are assigned based on syntactic criteria only: the label ACT is given to an element which has the syntactic function of subject but the semantic function of instrument. On the other hand, the other functors are assigned based on semantic criteria, as in these example sentences taken from Hajič et al. (2003): $H e_{A C T}$ made a canoe $P_{P A T}$ from the $\log _{O R I G}$ and $H e_{A C T}$ made a $\log _{P A T}$ into the canoe $e_{F F}$.

The corpora we will use for our analysis are annotated based on the analytical level of annotation of the PDT, so they refer to the theoretical framework of FGD. However, we will not follow the definition of valency given by FGD, since we extract verbal arguments from data annotated at the analytical level. Our definition of valency is thus purely syntactic and valency patterns (i.e. the sequences of the arguments for a verb) are defined in terms of the syntactic functions of the verb's arguments.

### 2.2.2.2 Subcategorization in Phrase Structure Grammar

After reviewing the DG view on valency in section 2.2.2.1, we here outline an alternative approach, referrable to the theoretical framework of PSG. We will contrast the different views of DG and PSG on syntactically representing verbal arguments; this choice seemed appropriate in the context of the present thesis because DG and PSG are the main models in treebank annotation and syntactic parsing, and represent the background to our work. On the one hand, a DG approach was chosen for describing the syntax of Latin in the corpora we used (see motivation in section 4.5.1) and for creating valency lexicons in lexicography (see iverview in section 4.3.1.1): a DG annotation is thus the starting point of our work on building an SC lexicon for Latin (chapter 4). On the other hand, an approach generally referrable to PSG was followed by the first systems for automatic acquisition of SC from corpora in computational linguistics (described in section 4.3.3): these systems constitute the methodological background to the design of our lexicon. We will thus illustrate the main difference between DG and PSG with respect to verbal arguments, and how this difference affects their views on SC. The generalization 'PSG' hides both subtle and profound variations, and the specific grammatical theories within this category are not treated in detail. For an overview of SC in various PSG theories (and not only), we refer to Korhonen (2002), while for a formal comparison between DG and PSG we refer to Fraser (1990).



Figure 2.2: PSG tree (left) and DG tree (right) for sentence (14) on page 20.

Figure 2.2 compares the phrase-structure tree and the dependency tree for sentence (14) on page 20. As shown in this example, in a DG tree every node corresponds to a word in the sentence and the edges represent the dependencies between the words: predicate ('Pred'), subject ('Sb'), Object ('Obj') and determiner ('Det') in figure 2.2.

On the other hand, in a PSG tree only the terminal nodes (leaves) coincide with the words of the sentence, and non-terminal symbols encode the constituents - such as noun phrases ' NP ' and verb phrases ' VP ' in figure 2.2 - or parts of speech - such as pronouns ('PR'), verbs ('V'), determiners ('DET'), nouns (' N ') in figure 2.2. As Mel'čuk (2003, p. 192) points out, the way the term head is used in DG does not coincide with the sense this term has in PSG: in sentence (14) the head of the NP the apple is apple in PSG, whereas its head (or governor) in DG is ate, which falls outside the noun phrase. ${ }^{3}$

The main property that distinguishes DG from PSG is the former's focus on dependency relations between lexical elements, rather than on phrasal constituents. Consequently, in DG verbal arguments are defined as special dependents of verbs, precisely those included in their valency. Unlike in a DG tree, in a PSG tree the syntactic functions of the constituents (such as subject, object, etc.) are implicitly encoded and can be inferred from the position of the nodes in the tree. Consequently, in PSG arguments are defined as special phrasal constituents inside the verbal phrase (VP), an exception being the subject, which is the phrasal node immediately under the sentence node.

These remarks show that DG is suited for morphologically rich languages with free word order like Latin. On the contrary, PSG is a suitable framework for a language like English, which is poorly inflected and shows fixed word order and few discontinuous constituents. This has an impact on systems for automatic acquisition of SC from texts. Since the first such systems were developed for English, they either used PSG-parsed data as input or they developed PSG parsers in order to detect verbal arguments: see for example the systems by Brent and Berwick (1991); Ushioda et al. (1993); Briscoe and Carroll (1997), which will be illustrated in section 4.3.3.1 on page 50. In these systems the order of constituents plays an important role in defining the syntactic functions and hence context features of VPs are heavily used to identify SC structures. For example, in Lenci et al. (2008) (see section 4.3.3.1, page 51) we worked on Italian and English and used the first post-verbal position as a heuristic rule to decide on argumenthood of a phrase. This approach would not be as efficient in the case of Latin, given the peculiarities of its morpho-syntax; a system for SC acquisition for Latin should thus exploit other properties, such as case and lexical features. Chapter 4 will face these issues and present our solution.

SC in Generative Grammar After highlighting the structurally different ways DG and PSG view verbal arguments, we will here give a general introduction to the role of SC in the Chomskian tradition, where PSG was formalised as part of a generative system. It is neither feasible nor relevant to this work to treat the place of SC in all theories belonging into the generative paradigm; instead, we will cover the definition of verbal arguments and some key concepts of the notion of SC in a number of selected generative approaches.

The syntactic tree to the left of figure 2.2 on page 22 is compatible with early generative analyses, where the phrase structure of sentences is expressed by rewriting rules paired with a lexicon containing the terminal symbols. For example, the left tree

[^4]in figure 2.2 can be generated by categorial rules such as $S \rightarrow$ NP VP (i.e. the initial symbol is composed by an NP preceding a VP), VP $\rightarrow$ V NP (i.e. the VP is composed of a verb followed by an NP), and lexical rules such as $\mathrm{V} \rightarrow$ ate (i.e. ate is a verbal form), DET $\rightarrow$ the (i.e. the is a determiner), $\mathrm{N} \rightarrow$ apple (i.e. apple is a noun). In this approach the arguments of a verb (except for the subject) are defined as those phrases included in the VP of the verb: for example the NP the apple in figure 2.2 is an argument of eat.

In order to simplify the form of phrase structure rules, the "Standard Theory" of Generative Grammar (Chomsky, 1965) introduced the notion of SC. In this theory context-free rewriting rules are distinguished into branching rules and SC rules: the former ones allow us to rewrite a category symbol as a string of either terminal symbols or non-terminal category symbols (such as NP $\rightarrow$ DET N); the latter ones "analyze a category into a complex symbol in terms of the frame in which this category appears" (Chomsky, 1965, p. 95), for example [+Animate $] \rightarrow[ \pm$ Human $]$. The new role played by the lexicon rests on the fact that not only does it contain the lexical items, but it also lists their idiosyncratic properties, such as declensional class for nouns or transitivity for verbs; these features are distinguished into selectional features and so-called strict SC features. For example, the entry for the noun sincerity would contain the features $+\mathrm{N},-$ Count (i.e. it is a non-countable noun), + Abstract (this is a so-called selectional feature), etc., while the entry for the verb frighten would contain $+\mathrm{V},+-\mathrm{NP}$ (this is a strict SC feature indicating that the verb is transitive), etc. (Chomsky, 1965, p. 107).

The X-bar theory (Chomsky, 1970) represents a further theoretical evolution from the Standard Theory. In the X-bar theory the arguments of a verb V are defined as those phrases which are sisters to the head V , while adjuncts are those phrases which are sisters to the single bar level $\mathrm{V}^{\prime}$. An illustration of this is provided in figure 2.3, where the argument the apple is an NP which is sister with V.


Figure 2.3: X-bar tree for sentence (14) on page 20.
Another important contribution of Chomsky (1970) is the so-called 'Lexicalist Hypothesis'. This hypothesis originated by the analysis of derived nominals (e.g. destroy/destruction). According to this hypothesis, similarities between deverbal NPs (e.g. the enemy's destruction of the city) and full sentences (e.g. the enemy destroyed the city) should be handled by lexical redundancy rules relating the verb with its derived nominal.

Later developments within Generative Grammar stressed this lexicalist approach, increasing the role of the lexicon in describing language. In Government and Binding

Theory (Chomsky, 1981) the lexical entries are said to contain semantic information about the meaning of words as well as their argument structure, represented in the so-called 'theta grid' (Haegeman 1996, pp. 49-72; Carnie 2007, pp. 221-226). Each predicate lexical item is associated with a theta grid containing the list of arguments it requires, paired with theta roles (i.e. bundles of thematic relations assigned to a particular argument). For example, in the sentence The teacher put the book on the table, the NP the teacher has two thematic relations (agent and source) and one corresponding theta role. The entry for put will contain three columns, one for each argument. Each column represents a theta role, with one or more thematic relation (Source/Agent, Theme and Goal, respectively), its category (NP, NP and PP, respectively) and an index which marks a theta role in an actual sentence (e.g. $i$ for Source/Agent, $j$ for Theme and $k$ for Goal, hence [The teacher] put [the book] ${ }_{j}$ [on the table] ${ }_{k}$ ). In order to stop the generation of ungrammatical sentences, the Theta Criterion imposes that each argument is assigned one and only one theta role and viceversa. Therefore the theta criterion, together with the rewriting rules, composes the computational component of the mind, which builds sentences and excludes ungrammatical ones. On the other hand, the lexicon records the "irregular" parts of the language such as the meaning of words, their PoS , their theta grid and other idiosyncratic features.

We can say that the theta criterion creates a structured interface between the syntactic rules and the lexicon through the theta roles. In section 2.2.2.1 we saw that DG faced a similar problem and found a solution which shares certain elements with the generative approach. Notably, at a semantic level, valency/argument structure features are allowed to interact with syntactic representations. Given the similarity of the solutions adopted by DG, PSG and Generative Grammar, the choice of theoretical framework and annotation should, for the purposes of studying SC from a computational perspective, be based on other criteria: the descriptive suitability of the framework for the language in question (e.g. the role of morphology in Latin syntax), what kind of computational tools are available for the framework, etc. These criteria led our choice of DG for building an SC lexicon for Latin verbs.

### 2.2.2.3 The lexical-semantic approach and the constructionist approach

In section 2.2.2.2 we saw that generative approaches postulate syntactic rules; in addition, they assume that the semantic features of verbs (and other lexical items) are contained in the lexicon, together with their argument structure features in terms of theta grids. In contrast with this, other approaches have postulated the existence of lexical rules mapping the semantic structure of the verbs themselves onto SC features: cf. Gropen et al. (1989); Levin (1993), among others. This approach, which we will call the 'lexical-semantic approach', is described here both for its theoretical relevance and for the specific computational applications it produced.

The lexical-semantic approach Jackendoff (1983) and Levin (1993) studied verb semantics and argument structure in light of the so-called 'syntax-semantics interface'. Their approach states that semantic properties of verbs help predict which syntactic alternations the verbs can occur with.

An alternation which has been widely studied is the English dative alternation, in which some verbs (give, send, etc.) allow for two possible syntactic realisations of theme and goal: object NP (theme) + PP (goal) and object NP (theme) + second object NP (goal), as in sentences (17) and (18), respectively.

He gave the book to her.
He gave her the book.
In her classification of English verbs, Levin notes that only certain semantic classes of verbs allow this alternation: verbs of future having (advance, allocate, offer, owe), send verbs (send, mail), verbs of throwing (kick, throw), etc.

This semantically-driven approach to syntax has been exploited by NLP research in several studies. Schulte im Walde (2003) and Lenci et al. (2008) use SC properties of verbs to cluster them into semantic classes. Similarly, the basic assumption of Korhonen (2002) and Alonso et al. (2007) is that the syntactic behaviour of a verb with respect to its arguments is largely determined by its meaning, and this is used in SC acquisition, as described in section 4.3.3.1, pages 51-51.

However, it must be pointed out that the interaction between semantics and syntax in argument realisation is not as predictable as such an approach would imply, since only certain aspects of verb semantics are responsible for argument structure. Levin and Rappaport Hovav (2005) discuss these issues and notice how difficult it is to isolate appropriate meaning components for characterizing specific verb classes, which often overlap. This has led many researchers to opt for a different, constructionist approach.

The constructionist approach Here we will present the constructionist approach, which is common to the different theories of Construction Grammar (Fillmore, 1985; Lakoff, 1987).

The lexical-semantic approach and the constructionist approch share the assumption that different SC behaviours correspond to different semantic properties of verbs. However, the constructionist approach excludes that the individual lexical items contribute to the whole range of semantic phenomena observed in languages. Hence, it claims the semantic contribution of the constructions themselves, independently of the particular lexical items. A construction is defined as a form-meaning correspondence (Goldberg, 1995, p. 1); this definition covers the whole range from words and idioms, to argument structure schemas and syntactic rules. Therefore, argument structures are constructions too, where the syntactic and the semantic-pragmatic components converge in a unique representation of the syntax-lexicon continuum.

Given its importance in lexicography, in what follows we will concentrate on one example of constructionist approach, referrable to the work by Charles Fillmore.

In Fillmore's Frame Semantics (Fillmore et al., 2003) semantic frames are introduced to account for the semantic content of constructions. A semantic frame conceptually describes a situation or an event typically evoked by a set of verbs, for example a commercial event, which is evoked from different perspectives by verbs like sell, buy, pay. A frame is composed by (core and non-core) frame elements, which can be thought of as semantic roles. For example, in the semantic frame 'Commerce scenario' the core elements are 'Buyer', 'Goods', 'Money' and 'Seller' and non-core elements are 'Manner', 'Means', 'Purpose', 'Rate', and 'Unit'. An example is given by the sentence: Ralemberg said he ${ }_{\text {Seller }}$ already had a buyer $_{\text {Buyer }}$ for the wine $_{\text {Goods }}$, a vintner living in Trinity. Frame Semantics is the theoretical framework of FrameNet (Baker et al., 1998), a lexical resource which describes a number of English predicates by their semantic frames, and correspondingly annotates corpus sentences (see section 4.3.1.3).

To conclude, we can say that section 2.2 .2 has shown how different linguistic theories deal with the issue of combining the contribution of syntactic and semantic factors to
verbs' argument structure. We started from Valency Theory within the DG approach, which defines verbs' valency in terms of the actants involved in the "process" expressed by the verb, to put it in Tesnière's terms; this has a direct effect on the syntactic realisation of these actants, given the syntactic and semantic nature of dependency. Then, we moved to PSG approaches and showed their structural differences with DG approaches; these differences have an impact on the syntactic representation of verbal arguments and, consequently, on how arguments can be extracted from sentences. As a further evolution of PSG grammars, the generative paradigm gave us the chance to explore the role of SC features in the lexicon and their link with the syntactic representation through the theta criterion. Finally, more semantically-oriented theories such as lexical-semantic approaches (Levin, 1993) and constructionist approaches (Goldberg, 1995) offered us the opportunity to investigate the connection between syntactic SC features and the verbs' semantics. In particular, the constructionist view moves the attention from the lexical items to constructions, responsible for the link between form and meaning, which is more than just a compositional sum of the two.

Considered the theoretical debate on these issues, the present thesis aims at showing a computational approach that deals with both syntactic and semantic aspects of verbs' argument structure. In addition, we will show how this task can be faced while considering the peculiarities of our language of study, Latin.

### 2.3 Latin verbal arguments

In section 2.2.2 we gave an introduction to verbal arguments and illustrated their general properties. However, argument structure phenomena are, to a large extent, language-specific. For this reason, some of the examples given for English in section 2.2.2 will be here contrasted with Latin examples, highlighting some morpho-syntactic and semantic properties of Latin verbal arguments (section 2.3.0.4). Finally, we will review previous lexicographic work by describing how SC and SP properties are represented in Latin dictionaries (section 2.3.0.5).

### 2.3.0.4 Properties

A central feature of verbal arguments regards their occurrence with verbal predicates: given the verb, this occurrence is more "expected" for arguments than for adjuncts. As we pointed out in section 2.2.2, this property has to do with the verb's semantics and the number of participants to the event described by the verb.

For example, sentence (19) exemplifies instances of arguments (murum, fossam) and adjuncts (interea, ea legione quam secum habebat) of the verb perduco 'bring through'.
(19) Interea ea legione quam secum
meanwhile that:ABL.F.SG legion:ABL.F.SG which:ACC.F.SG with oneself:ACC.SG
$\begin{array}{lll}\text { habebat } \ldots \text { murum } \ldots \text { fossamque } & \text { perducit } \\ \text { have:IND.IMPF.3SG } \ldots \text { wall:ACC.M.SG } \ldots \text {. and ditch:ACC.F.SG conduct:IND.PF.3SG }\end{array}$ (Caes.,
B. G., I)
'Meanwhile, he built a wall and a ditch with the legion he had with him'.
However, the realisation of the semantic participants is subject to pragmatic factors so that some of them may be omitted in certain contexts (optional arguments): cf. sentence (20), where the argument $e i$ 'to her' of the verb do 'give' is understood and omitted.
(20) Orabo ut mihi pallam reddat quam
ask:IND.FUT.1SG that me:DAT.SG mantle:ACC.F.SG give back:SBJV.PRS.3SG which:ACC.F.SG dudum dedi (Plaut., Men, 672)
a short time ago give:IND.PF.1SG
'I am going to ask her to give me back the mantle I gave her a short time ago'.
It is sometimes hard to distinguish optional arguments from adjuncts. In addition, this optionality, being realised at the morpho-syntactic level, appears to be different from language to language. For example, the event described by the Latin verb edo, just like the English verb eat, requires an eater and a thing eaten. However, it is not strictly necessary to specify both these elements if they are clear from the context or if this is not goal of the communication. For instance, the English sentence (5) on page 16 shows that the object of the verb eat (expressing the thing eaten in an active sentence) can be left out without affecting the grammaticality of the sentence nor the sense of the verb, i.e. it is optional. However, in English we cannot omit both the subject and the object of eat in a declarative main clause ( ${ }^{*}$ is eating). In general, the subject must be expressed in English declarative main clauses: this is what Herbst et al. (2004) call 'structural necessity'. Structural necessity for arguments is language-specific in a typological sense and applies at the clausal level. For pro-drop languages such as Latin and Italian the same necessity for subjects does not apply, as shown by sentences (21) and (22), where the number and person features of the verb inflection (third person singular) account for the subject of the verb.
(21) Mangia. eat:PRS.IND.3SG
'He/she eats.'
Edit.
eat:PRS.IND.3SG
'He/she eats.'
From this point of view, the Latin and Italian sentences (21) and (22) do not differ from sentences (23) and (24), where the verb is imperative: in all these sentences the subject is expressed by the verb morphology. In English, instead, the difference between the corresponding imperative sentence (Eat!) and the declarative sentence (5) (page 16) is more striking, since the former is formally subject-less, while the latter is not.

```
Mangiate!
eat:IMP.PRS.2PL
'Eat!'
```

Edite!
eat:IMP.PRS.2PL
'Eat!'
These remarks show that typological features of a language may affect the presence of verbal arguments in sentences. Hence, in Latin it is possible to omit any kind of argument, unlike in English. This broadens the gap between the number of realised arguments and the number of arguments theoretically required by a Latin verb; this affects all attempts to infer a verb's valency from a corpus rather than from an intuitionbased dictionary, which is what we tried to do in the SC lexicon we produced (chapter 4).

Now we will give an overview on valency classes for Latin verbs. In chapter 4 we
will adopt a more "empirical" definition of valency based on corpus occurrences, and discuss how to integrate the two definitions (section 4.8.1).

Valency classes The issue of argument structure of Latin verbs has been faced from different perspectives in Latin linguistics. Oniga (2007), for example, presents Latin syntax from the point of view of Generative Grammar, while other scholars refer to DG (Happ, 1976a; Pinkster, 1990; Griffe, 1995).

All these approaches use the term 'valency' to refer to the number of obligatory arguments for a verb. Consequently, any Latin verb belongs to one of more of the following classes: a) zero-valency (predicates referring to weather phenomena like pluit 'it rains' or tonat 'it thunders'), b) one-valency (predicates traditionally defined as intransitive, such as ambulo 'walk', where the only required argument is the subject, sentence (25)), c) two-valency (transitive predicates like laudo 'praise', sentence (26), intransitive verbs with an indirect object like memini, sentence (27), or copula verbs like sum 'be', sentence (28)), ${ }^{4}$ d) and three-valency (e.g. do 'give', sentence (29)).
(25) Caesar ambulat.

Caesar:NOM.M.SG walk:IND.PRS.3SG
'Caesar walks.'
Mulieres laudant deos.
woman:NOM.F.PL praise:IND.PRS.3PL god:ACC.M.PL
'The women praise the gods.'
Memini vivorum. remember:IND.PF.1SG living:GEN.M/N.PL 'I remember the living.'
Illi sunt proditores. they:NOM.M.PL be:IND.PRS.3PL traitor:NOM.M.PL 'They are traitors.'
Puer ei litteras dedit. child:NOM.M.SG he:DAT.M.SG letter:ACC.F.PL give:IND.PF.3SG
'The child gave him a letter.'
In order to show one of the main differences between Latin and English in argument realisation, in what follows we will underline the importance of morphology. Since for our SC lexicon we used dependency-annotated corpora where the argument/adjunct distinction is already encoded (see chapter 4), we will not insist on criteria for distinguishing arguments from adjuncts. Instead, we will be concerned with how to represent peculiar features of Latin verbal arguments in a lexicon.

The role of morphology The role played by morphology in Latin syntax is crucial. Morphological features help assign words to their PoS, as well as providing inflectional properties regarding their case, number, gender, etc. In particular, case inflection largely determines the syntactic role of a nominal element because each of the six cases is associated with a specific set of functions, although they sometimes formally overlap.

Traditional grammars for Latin make use of the term 'government' to refer to the fact that certain PoS classes (verbs, nouns and prepositions) select certain morphological

[^5]features of the elements with which they can combine, depending on their morphosyntactic functions. Traditional dictionaries use this term too, as we will see in section 2.3.0.5.

Government is connected with grammatical case: a verb governs a specific case if its complement must be inflected in that case for the structure to be grammatical. For example, the direct objects of the transitive verb laudo 'praise' must be in accusative as shown by sentence (30), while the verb pareo 'obey' governs the dative (sentence (31)).
(30) Eum laudabunt
he:ACC.M.SG praise:IND.FUT.3PL
'They will praise him.'
Paraverunt praecepto. obey:IND.PF.3PL order:DAT.N.SG
'They obeyed the order.'
Prepositions govern cases, too. For example, ad governs the accusative case, so that we can say ad septentrionem acc 'toward the north' but not ${ }^{*}$ ad septentrioni ${ }_{\text {dat }}$.

Government properties of verbs are related to their argument structure, since Latin case system - together with the system of prepositions - contributes to the syntactic realisation of the semantic roles expressed by verbal arguments. Several studies have dealt with the semantic relevance of cases and prepositions and their relation (Lehmann 1983, Carvalho 1983, Pinkster 1990, pp. 67-71, Carvalho 1996). An example of this interaction is provided by Latin preverbs, i.e. prefixes of verbal stems. Preverbs are morphological objects, but they affect the argument structure of the verb they are affixed to, as illustrated in chapter 6.

In what follows, we will outline the role played by cases and prepositions in differentiating the verbs belonging into the different valency classes. We aim at illustrating the range of possible realisations, although we will not give a comprehensive account of all possibilities. See section 4.5.2.2 for a complete description of these realisations in the annotation of the treebanks we used; see Pinkster (1990, pp. 39-52) for a more detailed description of the role played by cases and prepositions in Latin argument realisation.

Latin mainly follows the schema common to nominative-accusative languages (Rovai, 2008), where subjects of transitive and intransitive verbs receive the same case marking (nominative case), unlike direct objects of transitive verbs, which are marked with the accusative case, as shown by sentences (32) and (33).

Puer currit.
boy:NOM.M.SG run:IND.PRS.3SG
'The boy runs.'
Puer puellam videt.
boy:NOM.M.SG girl:ACC.F.SG see:IND.PRS.3SG
'The boy sees the girl.'

Valency one In one-argument predicates the only argument, if expressed, has the function of subject and agrees with the finite forms of the verb in person and number (and gender, when a participle is involved): cf. sentence (32).

Expressed subjects are mostly realised with the nominative case: cf. sentence (32). This realisation corresponds to the semantic role of agent for transitive verbs like laudo 'praise' or intransitive verbs like ambulo 'walk': cf. (30) (page 30) and (25) (page 29 ), respectively. It corresponds to Patient for the class of the so-called verba patiendi
like doleo 'suffer': cf. sentence (34). It can also correspond to other roles, such as experiencer (puer) in sentence (33).

```
Cor dolet.
heart:NOM.N.SG hurt:IND.PRS.3SG
'The heart hurts.'
```

Valency two For two-argument verbs, the second argument is usually an object. Its most common realisation is accusative (direct object): Pinkster (1990, p. 13) defines direct objects in general as "the constituent that becomes Subject in passive sentences with two- or three-place predicates": cf. puellam in sentence (33). Direct objects can also be expressed with subordinate clauses, such as interrogative clauses for verbs like dico 'say', e. g.: cf. ubi ero in sentence (36).

Other possible realisations of objects are what Pinkster (1990) calls 'complements' and we will call indirect objects: they are expressed by genitive (sentence (27) on page 29), dative (sentence (31) on page 30) and ablative (sentence (35)). These arguments express different semantic roles: for example, puellam in sentence (33), vivorum in (27) and praecepto in (31) are themes; gladio in (35) is an instrument.

Utitur gladio.
use:IND.PRS.3SG.DEP sword:ABL.N.SG
'He uses a sword'.
Tibi dicam ubi ero.
you:DAT.SG say:IND.FUT.1SG where be:IND.FUT.1SG
'I will tell you where I will be'.
Indirect objects can also be realised as PPs, for which there are several possibilities. Motion verbs, for example, require a locative object, which can be expressed in various ways: accusative (ibo Romam 'I will go to Rome'), ad/in 'to' + acc ( $a d$ forum 'to the forum'), adverb (ibi 'there'), subordinate clauses (ubi est 'where you are'), etc.

Another possibility for two-argument verbs involves copular verbs such as sum 'be', where the second argument is a predicate nominal: cf. sentence (28) on page 29.

Valency three For three-argument verbs, the third element is often expressed as a dative NP, as ei in sentence (29), which has the role of recipient of the event. Other possibilities include object complements (sentence (37)), various PPs (sentence (38)), subordinate clauses, etc.
(37) Fecerunt eum consulem.
make:IND.PF.3PL he:ACC.M.SG consul:ACC.M.SG
'They made him consul'.
Percussit eum cultello.
hit:IND.PF.3SG he:ACC.m.SG knife:ABL.M.SG
'He hit him with a knife'.
All we said here proves that morphological features involving case and lexical features involving the type of prepositions and conjunctions required for the realisation of verbal arguments crucially contribute to differentiate Latin verbs belonging to the same valency class. It is therefore essential to provide this information in an SC lexicon for Latin, as we will further discuss in section 4.4.1.

Word order The order of constituents in Latin sentences is considered to be relatively free. This variability mostly appears in Latin as the alternation between SOV, VSO, VOS, and SVO, depending on the relative position of the finite verb, its subject and its object. For a quantitative analysis of the distribution of SVO/SOV/VSO/VOS/OSV/OVS in the LDT, see Bamman and Crane (2006).

Moreover, non-projectivity phenonema are frequent in Latin, causing some constituents to be discontinuous, i. e. their continuity is interrupted by elements from other constituents.

In section 4.5.1, page 55 we show how word order and non-projectivity affect the syntactic annotation of Latin sentences. Here, we are interested in showing how these features affect verbs' argument structure. Sentence (39) contains the pattern VSO with two discontinuous constituents: the subject (ista gloria) and the direct object (mean canitiem).
(39) meam norit $\quad$ gloria $\quad$ canitiem $\quad$ (Prop,
that:NOM.F.SG my:ACC.F.SG know:SBJV.PF.3SG glory:NOM.F.SG white hair:ACC.F.SG
Carm, I, 8, 46)
'that glory would know my old age'

For languages such as English, context-based criteria have been proposed for indentifying arguments, as we hinted at in section 2.2.2.2 and will illustrate in section 4.3.3.1. These criteria are much less effective in Latin because of the phenomena just mentioned. For these reasons, it is important to take word order into account and record it in an SC lexicon, as we do: sections 4.4.2, 4.5.2.1 and A. 3 go into the details of this feature of our lexicon.

### 2.3.0.5 Lexicography

Valency theory had its first applications in lexicography, both for research-oriented monolingual dictionaries and for bilingual dictionaries used for teaching a foreign language, as we will show more in detail in section 4.3.1.1. However, the analysis of argument structure of predicates has always been part of traditional lexicography, even before Valency Theory; dictionary entries typically contain the arguments required by words in the form of agreement classes (e.g. gender), government and thematic roles.

Since we are interested in producing a lexical resource for Latin verbal arguments, we will give an overview of the role of SC and SPs in Latin lexicography. We are going to examine how some traditional resources for Latin deal with this information by looking at the Thesaurus Linguae Latinae (TLL) as well as monolingual and bilingual dictionaries. In addition, we will make some methodological remarks on the lexicographic approach used by the compilers of these resources; this will be contrasted with the approach followed by the valency lexicons and SC lexicons which will be analyzed in section 4.3. Both traditional Latin dictionaries and valency lexicons served as background to our work.

Forcellini (1940) This monolingual dictionary was first published in Padova in 1771. For each verbal entry, the SPs on the subject are represented using a fixed pattern with the preposition de 'about'; then morpho-syntactic constraints are given for each group of SPs. Not only the arguments of a verb are given, but also its most frequent adjuncts.

For example, in the entry for advenio 'arrive', the three subsenses for the first sense distinguish two groups of SPs, animate beings and inanimate things: "1.) de hominibus
aliisque animantibus, et 2.) interdum etiam de rebus inanimis". The first subsense receives further syntactic specifications regarding its occurrence alone or with adverbs: "A) absolute omnino, aut additis Adverbiis per tempus, in tempore, opportune etc.". Then, case features and the type of prepositions used in each subgroup of constructions are provided: for example, $a, a b, e x+$ ablative or adverbs for source complements "a) si agitur de termino a quo, cum Ablat. et praep. a, ab, ex aut Adv. unde etc. ...".

Lewis and Short (1879) This dictionary, completed in 1879, is now available online. ${ }^{5}$ It includes quotations from Classical and Late Latin and is based on second-hand material acquired from previous compilations.

Its entries follow a fixed structure: morphological and syntactic information on verbal arguments and adjuncts, subdivided by verb senses and subsenses. For example, the patterns for the first (main) sense of advenio 'arrive' are indicated as "constr. absol., with ad, in, or acc. ... with simple acc. Also with sup.". Each possibility is illustrated with examples.

Although it is not done systematically, SPs are expressed either by abstract semantic classes, or by the actual fillers of the arguments. For instance, for amo 'love', the SPs for its objects are given as: "to love a thing, to like, to be fond of, to find pleasure in, delight in: nomen, orationem, vultum, incessum alicujus amare.

TLL Even if only two thirds of it are complete, the TLL is considered as the lexicographic benchmark for Latin. ${ }^{6}$ The reference corpus of the TLL covers all Latin texts from the start until the second century A. D. and almost all inscriptions and texts from later eras until the fourth century A. D.

Each entry is provided with the senses and subsenses of the corresponding lemma. At the beginning of the entry some remarks on the word's semantics with respect to its synonyms are given. In addition, for each sense, morphological and semantic restrictions on verbal arguments are displayed, as well as some lexical fillers of the verb's adjuncts.

For example, in the entry for advenio 'arrive', the first sense of 'reaching a place' is defined with synonyms: "de loco (accedere, appropinquare, pervenire pedibus, navi, quo- modocumque)". The first subsense specifies the semantics of the agent role, restricted to humans (and animals): "homines (bestiae) adveniunt", adding examples and stylistic details on the spread of this subsense among Latin authors. Further semantic restrictions on the agent are given ('enemies' and 'army'): "de hostibus (in pugna supervenire, accurrere) ...exercitus vel duces exercituum", and morpho-syntactic properties of government (dative or ablative): "cum dat. vel abl". Moreover, the morpho-syntactic government of the first two senses are subdivided among the different locative complements. The explicit mention of case and preposition governed by the verb are usually provided, followed by examples, but sometimes examples only are given. In addition, SPs on the subject are displayed, either in terms of semantic classes (things, "res adveniunt") or as lexical examples (symptoms, "de morborum symptomatibus").

In the entry for the transitive verb amo 'love', a list of adverbs frequently occurring with this verb is provided after the SPs on its direct objects, showing an interest for both arguments and frequent adjuncts.

Generally speaking, morpho-syntactic features of verbal arguments and adjuncts are expressed either in explicit terms with the name of the case involved ("dat"), or

[^6]implicitly with the indefinite pronoun in that case ("alicui"), or sometimes directly with single words, phrases or larger fragments from texts. In a continuum between morphosyntactic realisation and lexical realisation, this information is offered to the reader in order to display the semantics of the verb and the variety of contexts it occurs in.

As to the structure of the thesaurus, no standard template for all lexical entries exists, but empirical norms have been established as the dictionary was growing. Moreover, the criteria for the partition of an entry in senses and subsenses are closely related to the textual material, so they change depending on the lemma.

Each sense is exemplified with text examples: only the entries of the less frequent lemmas record all texts containing those lemmas, while in the remaining entries a selection of passages is provided. The most general and common usages usually correspond to the lowest number of examples provided, especially in long lexical entries: as pointed out in the Introduction of the TLL, the number of quoted passages cannot be considered proportional to the material available in the corpus. The reason is that more emphasis is put on the diachronic evolution of the words and on their peculiarities. This approach to frequency issues differentiates the TLL from the lexicons automatically acquired from corpora, where frequency information is considered to be essential to illustrate the distribution of the word: cf. section 4.3.3.

Calonghi (1967) This bilingual dictionary addresses Italian students of Latin; therefore, its structure is concise and oriented towards explaining the meaning of words and their use.

For each sense of a verbal entry, semantic and morphological restrictions on the verb's arguments and adjuncts commonly occurring with the verb are given in Italian. The syntactic constructions are often associated with the Italian translation of the verb sense and with the equivalent construction in Italian. Morphological and syntactic information is sometimes merged together with semantic information.

For example, the first sense of advenio 'arrive' lists the SPs for its subjects (humans) and the possible morpho-syntactic patterns with the verb (absolute, i.e. with the subject only, with accusative, with dative of nouns referring to people, etc.): "di pers., assol., ...; col sempl. acc. ...col dat di pers. ...".

OLD The OLD, completed in 1982, uses a reference corpus of Latin literature until 200 A. D. which the compilers have gone through having the Oxford English Dictionary as a model.

It records, albeit in a non-systematic way, the patterns occurring with the verbs, with their case and prepositions constraints as well as semantic properties. For example, for advenio 'arrive', it shows the possible realisations of its locative complements: "1.(w. $a d$, in + acc. ) $\ldots$ (w. acc.) ...(w. adversus) ...(w. dat.) ...(w. advs) ...2. ...(w. adjs) ... (w. pple) ... (w. adv) ... (w. advl. acc.)". The SPs are associated to the verb senses and are indicated with English words, often showing an increasing specificity: e.g. "(of ships, vehicles) To arrive, put in. ... (of other things) to reach, be brought, arrive, ... (of possessions) to come into the hands (of) ... (of flowing water) ... (of wind) ...".

Summary The dictionaries and lexicons just reviewed represent both the general and the idiosyncratic aspects of valency/argument structure information in their entries. To do this, the lexicographer needs to access a textual collection; the degree of systematicity of this access can vary greatly. The TLL answers the need to exhaustively investigate
and categorize the texts available for a language by recording all or at least a large part of the words' occurrences. In other dictionaries examples are taken from the reference corpus with varying systematicity. In both cases the analysis of the corpus instances is done manually, hence a large amount of human resources, money and time is needed; more work is then necessary if the reference corpus is modified or enriched. In section 4.3.3 we will contrast these manual methods with automatic methods for recording SC information for verbs by accounting for all their instances in a corpus.

The degree of systematicity also affects the way the dictionaries are structured. The traditional lexical resources we have just reviewed are made by lexicographers to be consulted by readers and thus present inconsistencies, haphazard omissions or mistakes which are typically not systematic. In particular, the distinction between the different types of information (sense of the verb, morpho-syntactic realisations of complements, SPs) is subordinated to the goal of outlining the semantic nuances of the lemma. This helps the reader understand the meaning of the word and how it is used by the different authors. However, it makes it more difficult to retrieve SC information, both for purely linguistic studies and for computational uses.

### 2.4 Summary

This chapter has discussed the general background for the present thesis by focussing on its object of study, i.e. verb argument structure. We first gave an introduction to verbal arguments by contrasting them with adjuncts. We then illustrated how verbal arguments are treated in different linguistic theories, ranging from DG and PSG approaches to lexical-semantic and constructionist approaches. We focussed particularly on those theories which served as background for the valency lexicons described in chapter 4.

Then, we described some morpho-syntactic properties of verbal arguments in Latin, which is the topic of our work. Given the lexicographic focus of this thesis, after a theoretical introduction we surveyed various approaches to Latin verb argument structure in some traditional lexicographic resources, showing how the systematicity of these approaches can be improved for use in computational applications.

This thesis proposes an advance in the state of the art of computational resources for Latin by adapting computational methods developed for modern languages to Latin verbal arguments. In line with this plan, the next chapter will present the state of the art in computational linguistics for Latin.

## Chapter 3

## Computational resources and tools for Latin

In chapter 2 we gave an introduction to some theories on argument structure, which is the object of study of the present thesis. Since we aim at showing a new computational approach and resource to Latin verbs and their arguments, in this chapter we summarise the state of the art in computational linguistics for Latin, in order to show the material currently available to researchers in this field and how it can be enriched.

### 3.1 Introduction

As we saw in section 1.1, Latin is the language of the Index Thomisticus (Busa, 1980), the first electronic corpus, developed half a century ago. Nevertheless, research in computational linguistics applied to Latin has not kept up with the research carried out for modern languages and there is still a lot of work left to Latin computational linguists. Among the most widespread modern languages, a special attention is deserved to English. Computational linguistics studies of English are able to attract funding because of the impact this language has for present society. In the last decades, other European and Asian languages have been provided with more and more advanced linguistic resources, including both large electronic corpora (see definition in section 1.1.1.1) and tools for Natural Language Processing (NLP), whose applications involve various subfileds of Computational Linguistics and Artificial Intelligence.

In the rest of the chapter we will provide a picture of the main available computational resources for Latin, so that we will be able to motivate the "less-resourced" status of Latin stated in section 1.1.2. We will cover annotated corpora and NLP tools (section 3.2), as well as lexical databases (section 3.3).

### 3.2 Corpora and Natural Language Processing tools

In this section we give an overview of the main corpora available for Latin, with particular attention to annotated corpora. Section 3.2.1 introduces digital editions, section 3.2.2 describes manual and automatic morphological annotation, and finally section 3.2.3 tackles syntactic annotation.

### 3.2.1 Digital editions

The first step towards NLP for Latin consists of collecting the textual material in a format that can be handled by NLP tools. For this, it is necessary to digitise the immense amount of Latin texts produced throughout the long history of this language. This labour-intensive task has produced a large number of digital editions of Latin works, which will be briefly reviewed here. These editions can be browsed and searched through search engines: thanks to such tools, philologists, linguists and literary scholars can extract the occurrences of single word forms or sequences of word forms, together with their context (concordances) and frequency data on their distribution in the text.

Some of these editions were conceived to be used by philologists and therefore contain rich information on the tradition of the texts. As an example, we will consider the projects Musisque deoque and Poeti d'Italia in lingua latina tra medioevo e rinascimento, ${ }^{1}$ coordinated by scholars from the Italian universities of Padova, Trieste, Venezia Ca' Foscari and Verona. The aim of these projects is to collect the ditigal editions of Latin works by Latin authors from the Classical era and by Italian authors between Dante's birth and the first half of the sixteenth century. For each work the critical text is provided, including the codicological variants and, if possible, scholia and ancient glosses, as well as the Italian translation. On the website it is possible to search for the passages containing certain words (and all the variants obtained via inflection, compounding and derivation); it is also possible to impose contraints on the search such as the metre or the position in the line.

Other projects have chosen a specific edition rather then displaying the complete philological information. These digital editions are aimed at linguists and literary scholars as well as philologists, and can be searched by advanced search engines.

The first example we will consider in this brief oveview is the Index Thomisticus (Busa, 1980). It is available online ${ }^{2}$ and contains the best critical editions of the opera omnia by Thomas Aquinas (118 texts), together with works by other authors related to Thomas ( 61 texts), for a total of around 11 million words. The online version can be searched by word forms.

Another example is the Corpus Grammaticorum Latinorum (CGL), which contains the digitised version of the seven volumes by the grammarians of the Late Antiquity in the editions by Heinrich Keil (Leipzig, 1855-1880) or in more recent editions. This corpus is now available online ${ }^{3}$ and can be both browsed and searched by lexical forms.

Finally, the Library of Latin Texts (LLT) was jointly produced by the centre Cetedoc (CEntre de Traitement Electronique des DOCuments, at the Catholic University of Louvain-la-Neuve, headed by Paul Tombeur), by the publisher Brepols and by the centre CTLO (Centre Traditio Litterarum Occidentalium). The LLT corpus is available on CD-ROM. It is a database which is searchable by lexical forms and chronological eras, and includes the digital editions of Latin texts from the archaic age to the modern age. It contains 6 million words for the Classical era, over 20 million words for the patristic era, and over 600.000 words for the modern era, for a total of more than 50 million words.

[^7]
### 3.2.2 Morphological annotation of corpora

Once the texts have been digitised, they can be further analysed with various computer programs. The annotation the text has undergone determines the range of searches possible on it. For example, a lemmatised and morphologically annotated corpus displays the lemma and the morphological features of all its words; this allows the user to search not only the single forms in the text (e.g. all passages containing the form monui ${ }^{4}$ or puellae ${ }^{5}$ ), but also all occurrences of a given lemma (e.g. all passages where any form of the verb moneo is present, or where puellae is a singular genitive).

Now we are going to list some Latin resources which offer these possibilities of exploring the text, and stress the main characteristics of each of them from the point of view of their use in computational linguistics.

The annotation makes explicit some linguistic properties which are implicit in an unannotated text. Morphological properties can be included both with a manual annotation and with automatic methods. Since the corpora manually annotated at the morphological level also contain syntactic annotation, we will review them in section 3.2.3.1 and focus on automatic methods in section 3.2.2.1.

### 3.2.2.1 Automatic morphological annotation

The advantage of automatic annotation is that it allows savings of time and cost. In addition, automatic methods produce systematic errors which are easier to detect and correct than manual mistakes. Several software systems are available for producing a lemmatization and morphological analysis of Latin texts.

First, the Morpheus system ${ }^{6}$ (Crane, 1991) was developed in the Perseus Digital Library for Ancient Greek in 1985, and later extended to support Latin in 1996. The online digital library of the Perseus Project was founded in 1987 by the Department of Classics at the Tufts University and is supervised by Gregory Crane. It contains Latin texts for a total of more than 5 million words. Thanks to the analyzer Morpheus, the Perseus Digital Library is searchable by word forms and lemmas.

CHLT-LEMLAT ${ }^{7}$ (Passarotti, 2007a) was realised at the Institute of Computational Linguistics (ILC-CNR) in Pisa by Andrea Bozzi, Marco Passarotti and Paolo Ruffolo. This software provides all the possible morphological analyses of an input word.

Finally, another morphological analyzer is Words, ${ }^{8}$ designed by William Whitaker. Words is an elecronic dictionary for Latin containing 39000 lexical entries from different chronological eras. It gives the morphological analysis of an input word, together with its English translation. Rather than addressing Latin scholars, as Morpheus and LEMLAT do, Words aims at helping unexperienced Latin students and amateurs in simple translations. As its author acknowledges, the software behind Words automatically generates attested Latin words as well as "nonsenses".

### 3.2.3 Syntactic annotation of corpora

Syntactic annotation is the most advanced level of annotation available for Latin corpora. In this section we will survey manually syntactically annotated corpora (tree-

[^8]banks; section 3.2.3.1) and methods for automatic annotation of syntax (parsing; section 3.2.3.2).

### 3.2.3.1 Treebanks

A treebank is a collection of sentences manually annotated at the syntactic level and representend as syntactic trees. The rich information provided by treebanks can be exploited by traditional linguists for searching the corpora via syntactic criteria. In addition, treebanks have been widely used by computational linguists as a basis for developing automatic methods of syntactic annotation: see section 3.2.3.2.

A number of treebanks are available for different languages and follow different annotation schemas: for example, the Penn Treebank (Marcus et al., 1993) for English contains constituency trees, while the Prague Dependency Treebank (Hajič et al., 1999) for Czech contains dependency trees.

For what concerns Latin, three treebanks are available: the Index Thomisticus Treebank (IT-TB), the Latin Dependency Treebank (LDT) and the PROIEL Project Treebank. These treebanks are annotated following DG (see section 2.2.2.1): each word of a sentence is provided with a tag for its syntactic function and the dependency relation with its head.

IT-TB The IT-TB ${ }^{9}$ by the Catholic University of the Sacred Heart in Milan (Passarotti, 2007b) focusses on texts from the Index Thomisticus, introduced in section 3.2.1. The IT-TB originated from the project "Lessico Tomistico Biculturale" (LTB), supervised by Father Roberto Busa. It contains over 4,000 morphologically and syntactically annotated sentences (over 90,000 tokens) extracted from the concordances of the word forma in the Scriptum super Sententiis Magistri Petri Lombardi by Thomas Aquinas. The IT-TB can be browsed with the tool Netgraph, which allows for searches by lemma, form, morphological features, syntactic labels, and dependencies. See section A. 1 for more details on the format of the IT-TB.

LDT This treebank is being developed within the Perseus Project (Bamman and Crane, 2007). It shares the annotation manual with the IT-TB and consists of 53,143 tokens, extracted from works contained in the Perseus Digital Library (see section 3.2.2.1). It includes the following authors: Caesar (1,488 tokens), Cicero (6229), Jerome (8382), Ovid (4789), Petronius (12474), Propertius (4857), Sallust (12311), Vergil (2613). The annotation provides morphological and syntactic information, as will be described in section A.1.

PROIEL Project Treebank The third Latin treebank was started as part of the wider project PROIEL (Pragmatic Resources of Old Indo-European Languages). This multilingual project, supervised by Dag Haug at Oslo University, aims at producing a parallel corpus of the New Testment in Ancient Greek and of its translations in Latin, Gothic, Armenian and Old Church Slavonic. This parallel treebank is morphologically and syntactically annotated and contains 70,000 words.

CoLaMer Recently the project CoLaMer, collaboratively developed by the Universities of Köln and Regensburg, started to develop a fourth Latin treebank on texts of the Merovingian age.

[^9]
### 3.2.3.2 Parsing

Automatic syntactic annotation of corpora is perfomed by specific computer programs called parsers. A parser is rule-based if it exploits some manually constructed rules to parse a sentence; on the other hand, a statistical parser is trained on a treebank to "learn" its regulatities which can then be exploited when analyzing a new unannotated text. For example, if we train a parser on the IT-TB, it will learn the syntactic structures present in this corpus; after training it, we can use it to parse new portions of the Index Thomisticus or other corpora, which display similar stylistic features. As any other automatic method, parsing involves a percentage of error, which must be accounted for when using its output data. Automatically parsed data can be both used directly for further processing, or can be manually checked for a so-called semi-automatic annotation.

A first attempt at parsing Latin is described in Koch (1993), who applied an existing dependency parser by Covington (1990) to Latin. Koster (2005) describes a rule-based top-down chart parser automatically generated and developed from a grammar, as well as a lexicon built according to the formalism of the two-level AGFL grammar (Affix Grammar over a Finite Lattice; (Koster, 1991)).

McGillivray et al. (2009), Passarotti and Ruffolo (2009) and Passarotti and Dell'Orletta (2010) describe experiments with different statistical parsers for Latin. In particular, Passarotti and Dell'Orletta (2010) adapted the DeSR parser (Dependency Shift-Reduce; Attardi 2006) to better fit Latin sentences. They designed a feature model specific to Medieval Latin which aims at best capturing the regularities in Thomas' prose. Using a training set of 61,024 tokens ( 2,820 sentences), this caused an improvement of the previous accuracy rates, reaching $80.02 \%$ for LAS (Labeled Attachment Score, i.e. the percentage of tokens with correct head and relation label), $85.23 \%$ for UAS (Unlabeled Attachment Score, i.e. the percentage of tokens with the correct head), and $87.79 \%$ for LA (Label Accuracy, i. e. the percentage of tokens with the correct relation label; Buchholz and Marsi 2006). In addition, following the same approach, a PoS tagger was trained on the morphologically annotated data from the IT-TB, in order to automatically produce a morpho-syntactic disambiguation of the morphological lemmatization contained in the Index Thomisticus. Their preliminary experiments used the HMMbased HunPos tagger (Halácsy et al., 2007) and the IT-TB training set ( 61,024 tokens) and gave the following results: $96.75 \%$ of correctly disambiguated coarse-grained PoS + fine-grained PoS, and $89.90 \%$ if morphological features were also used.

Bamman and Crane (2008) report on experiments on PoS tagging of Classical Latin with the TreeTagger (Schmid, 1994) trained on a set of 47,000 tokens from LDT. Their accuracy rates in assigning the correct PoS are around $95 \%$.

LASLA Specific tools for morphosyntactic disambiguation and PoS tagging have also been developed by LASLA (Laboratoire d'Analyse Statistique des Langues Anciennes, University of Liège, headed by Joseph Denooz) for the annotation of their textual database, although they are not publicly available. The LASLA database consists of over one and a half million words and is provided with an online search module. ${ }^{10}$ This search engine allows for simple searches by lemma, bibliographic reference, or complete morphological analysis, as well as complex searches.

[^10]
### 3.3 Lexical databases

Several Latin dictionaries and lexicons are available online or on CD-ROM: the Lewis and Short dictionary provided by the Perseus Digital Library website, the Thesaurus Linguae Latinae from the Bayerische Akademie der Wissenschaften in Munich, the Thesaurus Formarum (TF-CILF) from the CTLO and the Neulateinische Wortliste made available by Johann Ramminger. ${ }^{11}$.

Here we will focus on the Latin WordNet (LWN) database (Minozzi, 2009), which is particularly important for its applications in semantic NLP tasks such as information extraction, data mining, word sense disambiguation and topic classification.

### 3.3.1 Latin WordNet

LWN is an ongoing project integrated within the existing MultiWordNet project, ${ }^{12}$ aimed at realising a large-scale multilingual computational lexicon based on WordNet (WN; Miller et al. 1990).

WN is a lexicon-oriented semantic network, started at Princeton University for the English language; in WN lexical items are organised into sets of synonyms called synsets and representing lexical concepts. The links among the synsets are defined by means of semantic and lexical relations, such as hyponymy, hyperonymy and meronymy. In particular, the synsets are linked to higher-level synsets (hypernyms) via tree structures until the top nodes. For example, the English noun bird belongs to five synsets; the first one ("warm-blooded egg-laying vertebrates characterized by feathers and forelimbs modified as wings") is linked to the top node [ENTITY, SOMETHING] via the intermediate hypernym nodes [vertebrate, craniate], [chordate], [animal, animate_being, beast, brute, creature, fauna], and [life_form, organism, being, living_thing].

Presently, the size of LWN is around 10,000 lemmas, with 9,000 synsets and 25,000 word senses. LWN represents a powerful resource for semantic processing of Latin, as proven by its use in our system for SPs described in chapter 5 .

### 3.4 Summary

In this chapter we gave an overview of existing computational resources for Latin, ranging from corpora to lexical databases and various NLP tools. This showed a gap in the state of the art which concerns both the size of the available annotated corpora and fact that the highest level of annotation reached is syntactic, leaving large space to research in computational semantics for Latin. These reasons motivate the contribution of our SC described in chapter 4 and our system for SP acquisition illustrated in chapter 5.

[^11]
## Chapter 4

## The subcategorization lexicon

### 4.1 Introduction

In this chapter we present a report of our work on building a computational lexical resource for Latin: a subcategorization (SC) lexicon for recording argument structure of Latin verbs. This resource was automatically extracted from two syntactically annotated collections of Latin texts (treebanks) and reflects the SC distribution in these corpora.

According to the definitions given in section 2.2, the valency of a verb refers to the set of arguments it requires, also called 'actants' or 'inner participants' (Bühler, 1934; Tesnière, 1959). These notions pertain to the so-called syntax-semantics interface in the sense specified in section 2.2.2.3. From a syntactic point of view, arguments are part of a verb's context and are structured in sequences called SC patterns. For example, a typical SC pattern found with the English verb give is $[\mathrm{NP}+\mathrm{V}+\mathrm{NP}+\mathrm{NP}]$ in a phrase structure representation and [Subject, Verb, Object, Object] in a dependency representation. Both these formalisms represent the pattern found in sentences like John gave me the book.

As we pointed out in section 2.3, for Latin verbal arguments we can talk about the morpho-syntax-semantics interface, given the vital role played by morphology in Latin argument realisation. For example, case features distinguish the two objects typically occurring with the verb do 'give', where the pattern ' X give Y to Z ' is expressed as ' $\mathrm{X}_{\text {nominative }}$ form_of_do $\mathrm{Y}_{\text {accusative }} \mathrm{Z}_{\text {dative }}$ '. For this reason, the lexicon we built contains morphological, syntactic and lexical information, as well as information about frequency. It lists which (lexically expressed) arguments occur with each verb in the corpora, along with their case, lemma, the verb's voice, and the number of times this co-occurrence is realised. In addition, the relative order of the arguments and the verb is recorded, as well as relevant intermediate elements such as prepositions, conjunctions, coordinations, and appositions. Finally, diachronic variables indicating the treebank (IT-TB or LDT) and the author allow us to use these data for diachronic studies, as described in chapter 7 .

In this chapter we first motivate our work from a computational linguistics perspective (section 4.2) and survey existing SC lexicons as well as previous systems for SC extraction and acquisition (section 4.3); then we describe our SC extraction system (sections 4.4 and 4.5) and the choices we made for representing this information in the lexicon (section 4.6). Some quantitative data from the lexicon are given (section
4.7 ), and finally a summary of this research is offered (section 4.8), suggesting future developments and indicating its use in the next chapters of the thesis.

### 4.2 Motivation

This section motivates the need for an SC lexicon for Latin, by summarising the main uses of an SC lexicon and its role in this thesis.

In section 1.1.2 we stressed the importance of creating computational resources for Latin and the role these resources play in favouring methodological advances in computational linguistics for Latin, as defined in chapter 1. In this perspective, the lexicon we introduce in this chapter contributes to advances in the state of the art of Latin (computational) linguistics in several ways.

First, the information provided can be used in corpus-based diachronic and synchronic studies on various aspects of Latin morpho-syntax: for example the argument structure of prefixed verbs (analysed in chapter 7), Latin word order (see section 4.7.3), the accusativus cum infinitivo construction compared with subordinate clauses introduced by quod (Bamman et al., 2008).

Second, SC lexicons are said to contribute to the expansion of linguistic resources by supporting the creation and enhancement of treebanks with regard to consistency of annotation: they provide annotators with essential information about the number and types of verbal arguments (Urešová, 2004). This means that our lexicon can help the development of both existing and new Latin treebanks.

Third, SC lexicons can be used for further research in Natural Language Processing (NLP) and computational semantics. Information about the syntactic roles and the lexical properties of verbal arguments helps the parsing of a sentence, as well as other NLP applications, such as word sense disambiguation, automatic verb classification and selectional preference (SP) acquisition (see, for example, Carroll et al. 1998, McCarthy 2001, Schulte im Walde 2003). In the present thesis, the lexicon is used as input to our SP acquisition system described in chapter 5 , which will add a lexical-semantic level of description to the verbal entries of the lexicon.

### 4.3 Background

In this section we will conduct a survey of existing SC lexicons, before demonstrating the gap in the state of the art which is filled by our work (section 4.4).

Section 2.2 gave an overview of how SC information is handled in traditional dictionaries for Latin. In the past decades several scholars belonging to different theoretical frameworks have compiled dictionaries and other lexical resources which contain (verbal) valency information for various languages. More recently, computational linguists have worked on automatically obtaining SC lexicons from corpora. Both these approaches present a higher degree of systematicity than traditional dictionaries, for example in the system of quotations from texts and references to corpus data. In addition, the lexicon entries follow common standards, for example by referring to a predefined set of categories for the argument properties. Such a coherent structure favours the implementation of NLP tools on these lexicons.

In this section we will describe a sample of manually compiled SC dictionaries (section 4.3.1), manually annotated corpora with argument structure information (section 4.3 .2 ), and automatically extracted SC lexicons (section 4.3.3). The terms used by
lexicographers to refer to the first group are 'lexicon' and 'dictionary', while the third group is usually referred to with the term 'lexicon'. In the rest of this thesis we will use the two terms as synonyms while talking about valency and SC lexicons: by these terms we will thus mean lexical resources recording some kind of information on words' valency and/or SC.

### 4.3.1 Intuition-based and corpus-based dictionaries

In the last decades the availability of large corpora and the ease with which they can be searched has had a great impact on the work of lexicographers. Compared with the past, they are able to access a significantly higher number of general and domainspecific occurrences of the lemmas of interest, which can be more accurately classified into different submeanings and accounted for from the point of view of their SC.

This section covers dictionaries which result from a manual and qualitative effort, where the corpus data are used as a source of examples to illustrate or derive the general behaviour described in the lexicon's entries. In section 4.3 .3 we will see that this approach differs from the approach used in lexicons automatically extracted from corpora, where the corpus data are systematically accounted for and constitute the basis from which a generalization is made via statistical and computational methods.

We will distinguish these manually built dictionaries on the basis of their theoretical framework: Valency Theory (section 4.3.1.1), Functional Generative Description (FGD; section 4.3.1.2) and Frame Semantics (section 4.3.1.3). We will concentrate on the way these dictionaries represent SC and SP information, since this is part of the background to our work illustrated in section 4.6 and chapter 5 , respectively.

### 4.3.1.1 Valency Theory

From a theoretical point of view, the first group of dictionaries we cover (Happ 1976a; Bianco 1996; Blumenthal and Rovere 1998; Herbst et al. 2004) refers to valency theory, which was introduced in section 2.2.2.1.

We will start from the lexicon contained in the Latin grammar by Happ (1976a), which is the only hand-made SC lexicon for Latin to our knowledge. It relies on a Dependency Grammar approach to Latin verbs' argument structure and focusses on the morphological aspects of verb valency. If compared with this first study, later lexicographic works based on Valency Theory (Bianco, 1996; Blumenthal and Rovere, 1998; Herbst et al., 2004) show an increased attention towards semantic, pragmatic and cognitive aspects of valency, as acknowledged by Helbig (1992) and Götz-Votteler (2007), among others. See section 2.2.2.1 for a discussion of these theoretical issues.

Happ (1976a) Happ's approach to verb valency stresses its morpho-syntactic aspects. Happ distinguishes arguments (Tesnière's 'actants', see section 2.2.2.1), which can be obligatory or facultative, from adjuncts (called 'circonstants'). He classifies arguments into seven types (Happ 1976a, p. 461; Happ 1976b): $A_{1}$ (nominative actants), $A_{2}$ (genitive actants), $A_{3}$ (dative actants), $A_{4}$ (accusative actants), $A_{5}$ (ablative actants), $A_{6}$ (actants which are prepositional objects), and $A_{7}$ (adverbial actants such as locative complements required by motion verbs).

Happ (1976a, pp. 480-565) reports a manually compiled list of Latin verbs, along with their valencies and some quantitative information extracted from a sample corpus of 800 verbal occurrences in Cicero's Orationes. Valency patterns are represented as sequences of symbols referring to the seven types of actants; morphological constraints
are provided on the case required by the arguments, and optionality is indicated by brackets.

For example, the sequence ' $A_{1} A_{4}\left(A_{6}\right)$ ' refers to the combination of two obligatory actants and a facultative one; their morphological realisations are expressed as ' $n-\operatorname{acc}-$ (prep+abl/acc)', meaning that the subject is nominative, the direct object is accusative and the optional argument is a PP followed by an ablative or accusative case. The three examples given for each of these three cases are: accipio 'take' $+\mathrm{acc}+a b$ 'from' +abl 'receive something from someone/something'; cognosco 'know' $+\mathrm{acc}+e x$ 'from' +abl 'know something/someone from something/someone; excito 'excite' + acc $+a d$ 'to' + acc 'excite someone to something'.

Bianco (1996) The Valenzlexikon Deutsch-Italienisch / Dizionario della valenza verbale (Bianco, 1996) is a bilingual German-Italian dictionary addressing Italian learners of German. The didactic aim coexists with a contrastive perspective on German and Italian verb semantics. The theoretical background to this work is Engel's theory and, in particular, his German-Romanian valency dictionary (Engel and Savin, 1983). Engel distinguishes adjuncts (Angaben in German) from 11 types of verbal arguments (Ergänzungen); the latter can be optional or obligatory.

Focussing on verbal valency, the lexical entries of Bianco's dictionary correspond to verbal senses via their Italian translations. For example, the German verb lieben has two Italian translations: amare 'love' and piacere 'like'. Each of these senses is assigned an SC pattern, represented as a sequence of figures, each of which refers to one of the 11 identified verbal arguments. These arguments are assigned syntactic features (subject, object, etc.), semantic features (temporal, locative, etc.), and case features. For example, ' 01 ' refers to the transitivity of the amare sense of the verb lieben (' 0 ' is the subject and ' 1 ' is the direct object). ' $0: \mathrm{UM}$ ' indicates that the subject is human. '!03, 0:., 3:UM a’ refers to the pattern for piacere which involves a subject and an object introduced by the preposition $a$ 'to', as in A Maria piacciono $i$ bei vestiti 'Maria likes nice clothes': this is an Italian peculiarity which is not present in German, hence is signaled by the exclamation mark. Finally, ' $3: \mathrm{UM} a^{\prime}$ ' shows at the same time the semantic restriction on the object (human) and the presence of the preposition $a$.

Blumenthal and Rovere (1998) The Italian-German bilingual dictionary by Blumenthal and Rovere (1998) ${ }^{1}$ consists of 1,729 verbal entries and 13,753 subentries, and contains a large number of example sentences from the 50 -milion-word Italian newspaper corpus Il Sole 24 Ore (issues 1989 and 1990), whose coverage of some specialised domains such as law and economics is said to be wide.

Each entry encodes the verb's translation and its SC patterns in the form of simple sequences. For example, one of the entries for cambiare 'change' corresponds to the German verb umziehen; it has the pattern 'N-V-N1', which means that it requires a subject and a direct object. Optional elements are distinguished from obligatory ones. The semantic part of the entry gives information on the verb's semantics (e.g. figurative) and on the arguments' semantics (SPs): for example, ' N : menschl.' for cambiare means that the subject is human.

The dictionary primarily addresses translators; due to the applied nature of the dictionary, the authors also record complements which might be considered as adjuncts, but which occur frequently with a given verb in the reference corpora.

[^12]Herbst et al. (2004) The first valency dictionary for English was published only recently and built by Herbst et al. (2004). It records the SC patterns of English verbs, nouns and adjectives. Unlike other dictionaries such as Bianco (1996), the authors clearly state the corpus-based nature of their work, based on examples from the Bank of English corpus ( 400 million words).

Each entry contains the different senses of the word and its valency as the number of required arguments in both active and passive sentences (for verbs), followed by at least one example from the corpus; in addition, the SC patterns are provided with their degree of optionality. The degree of optionality is encoded as obligatory, optional depending on the context, or optional tout court. Finally, semantic remarks on the meaning of the word in relation with the specific SC pattern are given, as well as the lexical fillers that frequently occur in these contexts. For valencies higher than two, the corpus frequencies of SC patterns are also given.

### 4.3.1.2 Functional Generative Description

Functional Generative Description (FGD; Sgall et al. 1986) was introduced in section 2.2.2.1. This theory is the basis for the annotation of the Prague Dependency Treebank (PDT; Böhmová et al. 2003; see page 20 ff .). The PDT-Vallex and the Vallex SC lexicons were built in the FGD framework, aiming to aid the annotation of the PDT. In this section we will describe both these lexicons.

PDT-Vallex and Vallex The PDT-Vallex lexicon contains 10,200 entries for Czech, half of which are verbs, and the rest are nouns and adjectives. The entries were manually created both during the annotation of the treebank and after 55,000 sentences were inserted in the treebank. Each sense of the entry lemma is associated with a valency frame consisting of a fixed number of slots; these slots correspond to functors expressing dependency relations - actor, patient, adressee, origin, effect - and are labelled with their optionality. Figure 4.1 shows an example of valency frames for a verb, a noun and an adjective, and is taken from Hajič et al. (2003). In each slot a functor is displayed ('ACT' for actor, 'PAT' for patient and 'ADDR' for addressee), together with its optionality degree ('obl' for obligatory) and its morphological features (cases). In the lexicon the valency frames are all equipped with an example sentence for illustration purposes.

| Word / | Valency Frame |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Sense | $\mathrm{Slot}_{1}$ | $\mathrm{Slot}_{2}$ | $\mathrm{Slot}_{3}$ |  | $\mathrm{Slot}_{n}$ |
| dát | $\mathrm{ACT}_{\text {obl }}$ (Nom) | $\mathrm{PAT}_{\text {obl }}(\mathrm{Acc})$ | $\mathrm{ADDR}_{\text {obl }}(\mathrm{Dat})$ |  |  |
| dopis | $\mathrm{ACT}_{\text {obl }}$ (Poss/Gen) | $\mathrm{ADDR}_{\text {obl }}$ (Dat) |  |  |  |
| plný | $\mathrm{PAT}_{\text {obl }}$ (Gen) |  |  |  |  |

Figure 4.1: Valency frames for the Czech verb dát 'give', the noun dopis 'a letter' and the adjective plny' 'full' in the FGD formalism.

Vallex (Lopatková, 2003) is another valency lexicon created within the FGD. It only contains 1450 verbs. The entries are structured as in PDT-Vallex, but they are the result of a more fine-grained research, adding a general semantic-cognitive category to each verbal entry corresponding to a verb sense.

### 4.3.1.3 Frame Semantics

Frame Semantics was introduced in section 2.2.2.3 and constitutes the background to FrameNet (Baker et al., 1998), a lexical database for English verbs. We chose to briefly illustrate the treatment of valency in FrameNet: for a full account of this issue (which falls outside the scope of this overview) we refer to Fillmore (2007).

FrameNet In FrameNet a verb corresponds to one or more annotated lexical units. Each of these lexical units is syntactically and semantically described in terms of the frame it bears. For example, the verb go has four lexical units corresponding to four frames: 'Becoming', 'Compatibility', 'Being_named' and 'Motion'. The valency patterns are thus sequencies of frame-specific semantic roles called frame elements (distinguished in core and non-core) within a semantic frame: for each frame element its syntactic realisations are provided. For example, the 'Motion' frame for go has 21 valency patterns. The first one is 'Carrier, Goal, Theme'. 'Theme' (i. e. "the entity that changes location") and 'Goal' (i.e. 'the location the Theme ends up in') are core elements, while 'Carrier' (i. e. 'the means of conveyance of the Theme') is non-core. The syntactic realisation for Theme is 'NP', those for Goal are 'INI' (Indefinite Null Instantiation) ${ }^{2}$ and 'DNI' (Definite Null Instantiation), ${ }^{3}$ and those for Carrier are ' $\mathrm{PP}[\mathrm{by}]$ ' and ' $\mathrm{PP}[\mathrm{in}]$ '.

FrameNet consists of 10,000 lexical units, corresponding to English word senses; for several of these entries annotated examples from the British National Corpus (BNC) and other corpora are provided.

We included FrameNet in the section on intuition-based and corpus-based dictionaries because of the way FrameNet is built: its developers first identify frames and annotate manually selected corpus sentences and then extract valency patterns and frame elements. Unlike the dictionaries described in sections 4.3.1.2 and 4.3.1.1, FrameNet can be defined as a lexical database; it can be searched by lemma as well as by frame or lexical unit at its url. ${ }^{4}$ Moreover, it contains a corpus annotated by frames which in turn can be browsed and searched. The format of the database is compatible with NLP tools, thus making it easier to use its data for further linguistic processing.

### 4.3.2 Manually annotated corpora

The interest towards corpus data, discussed for FrameNet in section 4.3.1.3, is even more marked in Propbank (Kingsbury and Palmer, 2002; Palmer et al., 2005), which is an annotated corpus.

PropBank Propbank adds a semantic layer of annotation to the syntactic layer of the Penn Treebank for English. The semantic content of the annotation in Propbank places it next to other projects, such as FrameNet (section 4.3.1.3) and the tectogrammatical

[^13]annotation of the PDT (section 2.2.2.1, page 21). However, the aim of this project is application-oriented rather than theoretical, aiming to facilitate the automatic extraction of data by NLP tools.

The focus of Propbank is on verbal predicates and their arguments, and the approach is theoretically neutral in the sense that the annotation labels are chosen independent from any particular theory of argument structure. The verbs are sense-disambiguated and their arguments are labeled as 'Arg0', 'Arg1' and 'Arg2', etc.; these labels have different meanings depending on the specific verb sense and hence do not correspond to abstract semantic roles such as the frame elements in Frame Semantics. For example, the first sense of the verb decontaminate ('clean completely') is annotated with the following arguments: 'Arg0: cleaner' and 'Arg1: dirty thing, now clean'. The verb-sense-specific nature of these labels is exemplified by the fact that $\operatorname{Arg} 0$, for instance, indicates the sender for the first sense of the verb send.

### 4.3.3 Automatically acquired lexicons

In section 2.2 .2 we noticed how lexicalist approaches to SC reduce the syntactic component and stress the lexical "idiosynchratic" nature of SC. Lexicalist theories aim at representing SC and valency properties of verbs in the lexicon rather than in the grammar. One way to represent this information is to make use of feature structures such as e.g. 'SUBCAT: DITRANS' or 'SUBCAT: NP,NP,PP' in the lexicon entry relative to a ditransitive verb like give. Lexicalist approaches are the (direct or indirect) background to most computational SC lexicons acquired from corpora, as we will see in this section. Given the aims of our work, we will focus on computational aspects of lexicography on SC rather than on the theoretical ones.

Over the past decades, research on automatic lexical acquisition has developed towards different directions: extraction of collocations (Smadja, 1993), SC (Brent, 1991; Ushioda et al., 1993; Briscoe and Carroll, 1997; Korhonen, 2002), SPs (Resnik, 1993; Ribas, 1994; Li and Abe, 1995), diathesis alternations (Lapata, 1999; McCarthy, 2001), and others.

As to SC, the need for a better account of domain-specific usages and the availability of relatively cheap, large corpora has made it possible to dynamically and automatically acquire or extract SC information to record in lexicons. This has established a close connection between the compilation of lexicons and the corpora themselves, since these lexicons, unlike manual ones, aim at systematically reflecting corpus evidence. Given the large size of today's corpora, this aim can be achieved via automatic methods, which have several advantages.

First, the automatic procedure makes these resources less costly than hand-crafted ones in terms of time, money and human resources. Second, these resources do not suffer from human errors such as omissions and inconsistencies, although they are prone to other types of (systematic) errors, which can be easily corrected once they have been identified: this makes them better suited for further automatic processing. Third, these lexicons are able to display quantitative information such as the frequency of SC patterns as attested in the original corpora. This is valuable both to lexicographers aiming at obtaining a full overview of the behaviour of words in corpora, and to computational linguists developing statistical systems that use these data such as parsers.

In this section we first review background work in the field of SC acquisition in computational linguistics (section 4.3.3.1), and then describe some automatically acquired lexicons (section 4.3.3.2).

### 4.3.3.1 SC acquisition

In this section we introduce previous systems for acquiring SC from corpora. We will distinguish 'extraction' from 'acquisition'. By the former term we mean the process of deriving SC information that is already present in the corpora, for example in the form of annotation labels. By the latter term we will mean the process of obtaining new SC information that is only implicitly present in the corpora, for instance separating arguments from adjuncts from a phrase-annotated corpus.

SC acquisition systems typically extract SC sequences in the form of SC frames (SCFs). ${ }^{5}$ For example 'NP, V, NP, PP-to' indicates a typical pattern occurring with the verb give (' X gives Y to Z '), where an NP subject ( X ) precedes the verb, which is followed by an NP object (Y) and a PP with the preposition to.

The first works on SC acquisition (Brent and Berwick, 1991; Ushioda et al., 1993; Briscoe and Carroll, 1997; Korhonen, 2002) focussed on English verbs and determined a set of SCFs; the acquisition systems, then, started from partially parsed data and aimed at identifying these same SCFs in corpora via rule-based and/or statistical computer programs.

Brent and Berwick (1991) Early research in SC acquisition is presented in Brent and Berwick (1991). Their system takes as input a PoS-tagged corpus - the 2.6 millionword Wall Street Journal corpus provided by the Penn Treebank project - and produces a partial list of the SCFs for each verb as output, with no frequency data. In particular, they concentrate on five SCFs: direct object, direct object + clause, direct object + infinitive, clause, and infinitive. The program uses a finite-state grammar to look for the easy cases where argument slots are filled by closed-class lexical items, mainly pronouns or proper names, rather than complete NPs. For example, rather than looking for the pattern 'V NP to V ', they look for ' V PRONOUN to V '. This way they avoid syntactic ambiguity, together with errors related to it; these restrictions, however, are so strict that most verb occurrences do not meet them, which is the major drawback of this system.

Ushioda et al. (1993) Ushioda et al. (1993) make use of a PoS-tagged version of the Wall Street Journal corpus (Marcus et al., 1993) and a finite-state NP parser which identifies six types of SCFs via some hand-written SCF extraction rules in the form of regular expressions.

They provide relative frequencies for the detected frames. In addition, they use a statistical method to increase the accuracy so that patterns of errors are learnt from training samples.

Briscoe and Carroll (1997) Briscoe and Carroll (1997)'s work differs from earlier systems in that a substantially larger set of SC types (163) is identified. This set has been a reference point for later work on SC acquisition for English verbs.

Briscoe and Carroll's system is able to build an SC dictionary, whose entries include the relative frequency of these SC classes. The goal is achieved by employing a PoS-tagger, a lemmatizer and a probabilistic left-to-right parser which yields complete though shallow parses of sentences. A patternset extractor then extracts SC patterns

[^14]for the verbs and is followed by a pattern classifier, which classifies the identified SC patterns as SCFs or rejects them as unclassifiable. The set of SCFs consists of around 30 manually identified SCFs, together with the SCF classes identified in two dictionaries for English. A patternsets evaluator, then, evaluates the verb patternsets and constructs potential SC entries; it finally filters them by binomial hypothesis testing on frequency data.

Korhonen (2002) Briscoe and Carroll's system is improved by Korhonen (2002), who proposes an acquisition method based on semantic properties of verbs. Korhonen assumes that semantically similar verbs are also similar from an SC point of view, by exploiting Levin (1993)'s taxonomy of English verbs.

All the previous systems were built for English; for languages other than English, a repertoire of SCFs is not available. For this reason, SC acquisition for these languages aim at 'discovering' frames as they are extracted from corpora: such a "dynamic" definition of SCF leaves room for including typical verbal modifiers in addition to canonical verbal arguments.

Basili et al. (1997) Basili et al. (1997) work on domain-specific corpora for Italian. They first tokenize, lemmatize and parse the input corpus; then they use a supervised clustering method to indentify the different senses of a verb and associate the corresponding SCFs.

Zeman and Sarkar (2000b) Zeman and Sarkar (2000b) start from the PDT (see section 2.2.2.1, page 21) and use a machine learning technique to acquire SCFs for Czech. Given the dependency annotation of the PDT, they can extract the set of all dependents of a verb in a tree ('observed frame'). The verb's SCF is a subset of this set containing arguments only; it is found by collecting all subsets of the observed frame and excluding large infrequent subsets (which are likely to contain adjuncts) as well as small infrequent subsets (which are likely to miss some arguments). The remaining observed frames are scanned by subsets and the best subsets (SCFs) are selected based on statistical tests.

Alonso et al. (2007) Alonso et al. (2007) deal with Spanish and use a syntactically and semantically annotated corpus - the SenSem Corpus (Castellón et al., 2006) - to form equivalence classes of verbs based on their SCFs. The rich information contained in the SenSem corpus includes senses of the 250 most frequent Spanish verbs, SCFs and semantic roles of the verbs' arguments; following the syntax-semantics interface hypothesis (see section 2.2.2.3), this information is exploited to cluster verb senses on the basis of the realisations of their arguments in the corpus sentences. The final task is to assign SCFs to unseen verbs.

Lenci et al. (2008) In Lenci et al. (2008) we worked on Italian, a language for which a predefined set of SCFs is not available. This justified a 'tabula rasa approach' which does not require any previous knowledge about SCFs and where SCFs are defined as they are found in the corpus. This strategy has several advantages, allowing us to identify lexically selected modifiers, as well as arguments, which are likely to be of useful for NLP-based applications.

In this work we used a chunked corpus (LE PAROLE corpus; Goggi 1997), where the sentences are syntactically analysed into non-recoursive phrases. The SC acquisition
proceeds by extracting verb local contexts from the chunked corpus and concentrates on right contexts, based on the assumption that this is where verbal arguments tend to gather, given the relatively fixed order of sentence elements in Italian. These local contexts are then reduced in length to potentially subcategorized slots which are part of the SCF and contain chunks that occur in the first slot to the right of the verb: this follows the heuristic rule that the first slot to the right of the verb is the position held by arguments; the potentially subcategorized slots are further selected with a frequency threshold to obtain full SCFs.

Lenci et al. (2010) In Lenci et al. (2010) a distributional method for extracting SC from the large automatically parsed corpus of La Repubblica is presented. In addition, SPs are extracted based on the statistical measure of Simple Log Likelihood. The results are included in an ongoing project for building an Italian version of FrameNet.

### 4.3.3.2 Lexicons

Most of the previous systems for SC acquisition aimed at building an SC lexicon to be employed for lexicographic purposes or in NLP applications. These two directions have also led to two approaches to evaluation: on the one hand the acquired SCFs were evaluated by manual inspection (Brent and Berwick, 1991; Ushioda et al., 1993; Korhonen, 2002); on the other hand, the evaluation was performed on a specific task: parsing in Briscoe and Carroll (1997), verb clustering in Lenci et al. (2008).

In this section we will describe a sample of these automatically acquired lexicons, since they are part of the background to the design of our lexicon.

Valex Valex (Korhonen et al., 2006) was obtained by using the system developed by Briscoe and Carroll (1997) and Korhonen (2002). This lexicon was extracted from five corpora and from the web and contains SCFs for 6397 English verbs.

Each entry corresponds to a verb and is subdivided into SCFs, identified by the corresponding number in Briscoe and Carroll's list (Briscoe and Carroll, 1997). Then, the frequency of the verb and the SCF in the corpus is given, together with the lexical fillers of the arguments.

Table 4.1 shows the five most frequent SCFs for the verb go as they appear in Valex, indicated after the label 'SUBCAT', then the class of the SCF in Briscoe and Carroll's list ('SUBCAT'), the relative frequency ('RELFREQ') and the absolute frequency ('FREQCNT').

SUBCAT (VSUBCAT NONE) :CLASSES (22) RELFREQ 0.583292 :FREQCNT 3498
SUBCAT (VSUBCAT PP) :CLASSES (87) RELFREQ 0.117225 :FREQCNT 703
SUBCAT (VSUBCAT NP) :CLASSES (24) RELFREQ 0.097549 :FREQCNT 585
SUBCAT (VSUBCAT AP) :CLASSES (1) RELFREQ 0.073203 :FREQCNT 439
SUBCAT (VSUBCAT VPINF) :CLASSES (112) RELFREQ 0.042188 :FREQCNT 253
Table 4.1: Five most frequent SCFs for go in Valex.

LexSchem LexSchem (Messiant et al., 2008) was produced with the same approach used for Valex and contains SC for 3297 French verbs.

Table 4.2 shows the SCFs ('Cadre de sous-catégorisation') for the verb aller 'go', followed by the relative frequency ('Nombre d'occurences') and the absolute frequency ('Fréquence relative').

| Cadre de sous-catégorisation | Nombre d'occurences | Fréquence relative |
| :--- | :--- | :--- |
| INTRANS | 20156 | 0.135 |
| SINF | 117937 | 0.791 |
| SN | 11089 | 0.074 |

Table 4.2: SCFs for aller 'go' in LexSchem. The first column contains the SCF, the second the absolute frequency and the third the relative frequency.

Bamman and Crane (2008) This is the only automatically acquired SC lexicon for Latin. Bamman and Crane (2008) illustrate a "dynamic lexicon" automatically extracted from the 3.5 million-word corpus based on the Perseus Digital Library. The corpus was processed by an automatic morphological tagger and a statistical syntactic parser using the LDT as training corpus. The authors' procedure reduced the noise caused by the automatic pre-processing of the data through the extraction of the most common arguments (SCFs) and their most common lexical fillers. The dynamic nature of the lexicon refers to the fact that it can be updated as the treebank annotation proceeds.

The lexicon contains qualitative and quantitative information on SC patterns and their fillers as they are used by each author of the corpus. Each entry records the word's sense given by the lemmatization based on the dictionary Lewis and Short (1879). In addition, the most common SCFs for the Latin verbs are obtained thanks to the automatic dependency parsing, which identifies the words' arguments. Finally, SPs are represented in the form of most common lexical preferences of verbs on their arguments, as explained more in detail in section 5.2.2.3 on page 91 .

### 4.4 Our contribution

Before giving an overview of our system for creating an SC lexicon for Latin (section 4.4.2), we will outline our contribution and how our lexicon has the desirable features of a computational SC lexicon (section 4.4.1).

### 4.4.1 A new resource for Latin

We propose a computational lexical resource that helps the advance of the state of the art in computational linguistics for Latin. More specifically, we introduce a method for extracting an SC lexicon for Latin verbs from two treebanks; for a nearlier version of the lexicon see McGillivray and Passarotti (2009). The information on SPs will be added via the method illustrated in chapter 5.

The SC extraction proceeds from the data provided by the Latin Dependency Treebank (LDT; Bamman and Crane 2007) and the Index Thomisticus Treebank (IT-TB; Passarotti 2007b). For a description of the LDT and the IT-TB treebanks see section 3.2.3.1.

Our method is based on previous work in automatic SC acquisition and manual construction of SC lexicons (section 4.3); it was adapted to the pecularities of Latin
morpho-syntax described in section 2.3.0.4 and inspired by the long lexicographic tradition in building Latin dictionaries which we discussed in section 2.3.0.5. For these reasons, our work places itself at the interface between computational lexicography and lexical acquisition in computational linguistics on the one hand, and Latin syntax and lexicography on the other.

Here, we want to briefly list the desirable features of an SC lexicon that pertain to our lexicon. First, each entry corresponds to a verb occurring in the corpora and there are around 20,000 of them, which is a considerable figure if we compare it with the other resources available for Latin (apart from collections of raw text). The lexicon entries represent all the SCFs found with the verb, together with their frequencies. In addition, given the importance of morphology in Latin verb argument structure (section 2.3.0.4), morphological attributes such as the voice of the verb, as well as the PoS, the case (for nouns, pronouns and adjectives) and mood (for verbal forms) of the arguments are recorded. Lexical restrictions on prepositions, as well as coordinating and apposing elements introducing the arguments are specified as well. As to semantic content, in chapter 5 we will describe a system for acquiring SPs of verbs which can be included in the lexicon. Finally, the IT-TB part of the lexicon is available through a graphical interface at http://itreebank.marginalia.it/itvalex, among the activies of the CIRCSE research center at the Catholic University of Milan. ${ }^{6}$ The interface, which is under development at the moment of writing, will make the results of queries on the lexicon available for download: this is useful - among other things - for quantitative linguistic research of the type reported on in chapter 7 .

### 4.4.2 Overview of the system

Here we summarise the main steps we followed while extracting SC from the corpora and representing it in the lexicon. The system will be described in more detail in sections 4.5 and 4.6 .

1. The first step consists in the automatic extraction of verbal arguments from the treebanks, along with their syntactic roles, and their morphological and lexical features. The arguments are retrieved using database queries, exploiting the fact that in both treebanks they are annotated using specific tags: 'Sb' for subjects, 'Obj' for direct and indirect objects, 'Pnom' for predicate nominals, and 'Ocomp' for object complements (see Bamman et al. 2007b). A non trivial problem regards extracting possible "intermediate" nodes occurring between the verb and its arguments, such as prepositions or coordinations.
2. After acquiring these data, the lexicon is ready to be built. For what concerns the formal representation of SC structures, they are recorded in terms of SC frames; these frames contain the syntactic functions of the verbal arguments making up the frame in the order with which they appear in the text. Being Latin a free word order language, we also defined SC classes as syntactic patterns where the SC elements are represented disregarding their linear order in the sentence. For instance, one of the possible frames for the verb compono ('to link') in the IT-TB is ' $\mathrm{Sb}[\mathrm{nom}]+\mathrm{V}+\mathrm{Obj}[\mathrm{acc}]+($ cum $) \mathrm{Obj}[\mathrm{abl}]$ ', corresponding to the occurrences of this verb preceded by a nominative subject and followed by an accusative object and an ablative object depending on a preposition, as in this sentence: intellectus

[^15]componit privationem cum subiecto ('the intellect links privation to a subject'). The corresponding SC class is '(cum)Obj[abl], Obj[acc], $\mathrm{Sb}[n o m]$ '.

### 4.5 Subcategorization extraction

This section describes our system for extracting SC from the two Latin treebanks we used: the LDT and the IT-TB. Section 4.5.1 illustrates the specific annotation schema adopted for the LDT and the IT-TB, which is based on Dependency Grammar (DG; see section 2.2.2.1). Section A. 2 explains how this particular annotation style affected the database queries we designed for the SC extraction. Although the two treebanks focus on different time periods, they were annotated following the same annotation style and format. For this reason, it was possible to run the same database queries on both corpora (with some adjustments), thus obtaining two comparable SC lexicons and allowing linguistic comparisons between the different eras represented (see, for example, the analysis on Latin preverbs reported in chapter 7).

### 4.5.1 Dependency annotation for Latin

In section 2.2.2.1 we outlined the main theoretical assumptions behind the DG paradigm. Since their start in 2005, both the LDT and the IT-TB independently chose Dependency Grammar (DG) as the most suitable theoretical framework for describing the syntax of Latin sentences. Here we will go through some general features of DG theories applied to treebank annotation and describe the DG annotation systems shared by the LDT and the IT-TB.

Dependency-based theories have gained popularity in the last years, both in natural language parsing and in treebank development. Regarding parsing procedures, (Collins, 1996) shows that using dependency probabilities between heads in the parse tree improves the performance of a statistical parser. Moreover, Lin (1995) and Carroll et al. (1998) propose two dependency-based evaluation systems for parsers, explaining the importance of such an approach for semantic interpretation. Regarding treebank development, over the past decade DG was adopted by a number of treebank projects, such as the ALPINO treebank for Dutch (Van der Beek et al., 2002), the PDT for Czech (see page 21), and the Danish Dependency Treebank (Kromann, 2003).

The reasons behind the adoption of DG in building treebanks are various. Generally, DG annotation is considered easier to learn by annotators than parsing rules, as dependencies are less heavily affected by the grammar formalism than parsing structures. A more language-specific reason for adopting DG in treebanks is that DG is generally considered to be suitable for inflectionally rich languages with free word order because it can handle non-projectivity better. Some interrelated features of the Latin language - such as rich morphology, non-projectivity, and free word order - explain the adoption of DG in building Latin treebanks.

Free word order In section 2.3.0.4 we briefly introduced free word order in Latin, i. e. the high variability in the position of the constituents in Latin sentences. Such a high variability is made possible by Latin's rich morphology, which theoretically allows all of the semantically equivalent sentences listed in table 4.3 , obtained by changing the order of the words in sentence (1).
(1) Tenet ille immania saxa (Vergil, Aeneis, I, 139)
possess:IND.PRS.3SG he:NOM.M.SG enormous:ACC.N.PL rock:ACC.N.PL
'He possesses the enormous rocks'


Table 4.3: Sentences obtained from sentence (1) on page 55 by changing the order of the words, and corresponding DG representations.

Non-projectivity Free word order is also related to non-projectivity, that is the possibility of constituents in a sentence to be broken up by elements belonging to other constituents. Non-projectivity has a clear graphical interpretation in dependency structures, as shown in the first column of table 4.3. ${ }^{7}$ The first two rows display two cases of projectivity (non-overlapping arcs), whereas the last row represents a nonprojective diagram (overlapping arcs), because of the position of the phrase immania saxa between the subject ille and the predicate tenet. Such cases of non-projectivity can be easily represented by DG trees, as no movements are required to record the actual order of the words in the sentences. As an illustration of this, see the second

[^16]column in table 4.3, where the linear order of the sentence is kept when reading the trees from left to right.

Non-projectivity is frequently attested in Latin texts, which can display quite complex discontinuous constituents, as illustrated in figure 4.2, referring to sentence (2).

```
    which:ACC.M.SG to end:ACC.M/F.SG oneself:ACC.SG unbridled:NOM.F.SG throw:IND.FUT.3SG
    audacia? (Cicero, In Catilinam, speech I, chapter 1)
    audacity:NOM.F.SG
'To what end will your unbridled audacity throw itself?'
```



Figure 4.2: A DG representation of sentence (2).

Table 4.4 contains the number of non-projective tokens in the LDT and in the ITTB over the total number of tokens. A Pearson's $\chi^{2}$ test of indipendence (see section C. 2 for a description of this test) on these data shows that, even though the number of non-projective tokens is significantly higher for the LDT, the size of this difference is very small $\left(\chi_{(1)}^{2}=636.5696, p<2.2 e-16\right.$, effect $\left.\operatorname{size} \varphi=0.08\right) .{ }^{8}$

| corpus | projective | tokens <br> non-projective | total | non-projectivity rate |
| :--- | :---: | :---: | :---: | ---: |
| LDT | 49,607 | 3,536 | 53,143 | $6.65 \%$ |
| IT | 53,057 | 1,821 | 54,878 | $3.31 \%$ |

Table 4.4: Non-projectivity in the LDT and in the IT-TB.

### 4.5.2 The annotation of the LDT and the IT-TB

In this section we will focus on the annotation lines followed by the LDT and the IT-TB, as they help illustrate how our SC lexicon was obtained. We will first give an overview on the annotation through an example (section 4.5.2.1), then present the annotation guidelines of the two treebanks (section 4.5.2.2). For the format of the two treebanks, see section A.1.

### 4.5.2.1 Overview

The annotation of the LDT and the IT-TB assigns a rooted tree to each sentence in the corpora; the nodes of the tree correspond to the tokens of the sentence (including punctuation marks) and the edges represent either the binary syntactic dependency relations between node pairs or other phenomena such as coordination, apposition, punctuation, etc. (labelled with functional labels). The tree also records the original

[^17]linear order of the elements in the sentence. Figure 4.3 presents the dependency tree of sentence (3) from the IT-TB.
(3) Praedicatum enim habet fationem formae. (Th., Predicate:NOM.N.SG indeed have:IND.PRS.3SG account:ACC.F.SG form:GEN.F.SG Sent. P. L., I, Q. 1, Art. 4, Arg. 1, 6-1.7-8)
'The predicate indeed accounts for the form.'


Figure 4.3: The dependency tree of sentence (3) from the IT-TB.

In this sentence, there is a subject ('Sb') relation between the word praedicatum and the word habet, an object ('Obj') relation between rationem and habet, and an attribute ('Atr') relation between formae and rationem; moreover, the finite verb ('Pred_Co') depends on the conjunction enim, which in turn depends on the top node, labelled as 'AuxS' and representing the whole sentence. The punctuation mark (tagged as 'AuxK') is also dependent on the top node.

### 4.5.2.2 Syntactic annotation guidelines

In this section we will describe the main features of the annotation guidelines of the LDT and the IT-TB. We will focus our attention on the syntactic functions and phenonema we accounted for while building the SC lexicon.

Since they share the DG framework of annotation, the LDT and the IT-TB have collaborated since the beginning of their respective projects, which were the first ones for Latin. As no guidelines had been designed for syntactically annotating Latin corpora at that time, these two projects decided to follow the guidelines for the so-called 'analytical layer' of annotation of the PDT (Hajič et al. 1999; see section 4.3.1, page 47) and adapt them to Latin syntax following Pinkster's grammar (Pinkster, 1990).

Table 4.5 lists all the syntactic tags currently in use in the LDT and the IT-TB. ${ }^{9}$ In

[^18]| Pred | predicate |
| :---: | :---: |
| Sb | subject |
| Obj | object |
| Atr | attributive |
| Adv | adverbial |
| Atv/AtvV | complement |
| PNom | predicate nominal |
| OComp | object complement |
| Coord | coordinator |
| Apos | apposing element |
| AuxP | preposition |
| AuxC | conjunction |
| AuxR | reflexive passive |
| AuxV | auxiliary verb |
| AuxX | commas |
| AuxG | bracketing punctuation |
| AuxK | terminal punctuation |
| AuxY | sentence adverbials |
| AuxZ | emphasising particles |
| AuxS | root of the tree |
| ExD | ellipsis |

Table 4.5: Complete Latin tagset adopted by the LDT and the IT-TB.
addition, all the tags (except for 'AuxP' and 'AuxC') can be appended with one of the '_Co', '_Ap', and '_Pa' suffixes, like in the PDT, depending on whether the nodes occur respectively in a coordination, an appositive construction, or a parenthetical statement. See page 67 for examples of coordination and apposition.

In the next sections we will list the annotation of verbal arguments in the LDT and IT-TB, exemplifying some of their morpho-syntactic realisations in the SC lexicon we built. We will also mention other syntactic tags that are relevant for understanding the entries of the lexicon. For a detailed illustration of the linguistic annotation, see the guidelines in Bamman et al. (2007a). ${ }^{10}$

We will refer to the following glossed sentences from Caesar, B. G., II, 1-2, whose trees are shown in figures $4.4,4.5,4.6,4.7,4.8$, and $4.9 .{ }^{11}$

| (4) Cum esset | Caesar | in citeriore | Gallia | in |
| :--- | :--- | :--- | :--- | :--- |
| when be:SBJV.IMPF.3SG | Caesar:NOM.M.SG in Hither:ABL.F.SG Gaul:ABL.F.SG in |  |  |  |

[^19]Romanum coniurare obsides -que inter se dare.

Roman:ACC.M.SG conspire:INF.PRS hostage:ACC.M.PL and between oneself:ACC.SG give:INF.PRS 'While Caesar was winter quarters in Hither Gaul, as we have shown above, frequent rumours were brought to him, and he was also informed by letters from Labienus that all the Belgae, whom we have alread described as a third of Gaul, were conspiring against the Roman people and giving hostages each to other'


Figure 4.4: The dependency tree of sentence (4) from the LDT.

> (5) Ipse, cum primum pabuli copia esse
> He:NOM.M.SG, when first:ADV forage:GEN.N.SG abundance:NOM.F.SG be:INF.PRS inciperet, ad exercitum venit
> begin:SBJV.IMPF.3SG, to army:ACC.M.SG come:IND.PF.3SG
> 'He himself, as soon as there started to be a supply of forage, came to the army'
(6) Dat negotium Senonibus reliquis -que Gallis
give:IND.PRS.3SG task:ACC.N.SG Senones:DAT.m.PL other:DAT.m.PL and Galli:DAT.M.PL


Figure 4.5: The dependency tree of sentence (5) from the LDT.

| qui | finitimi | Belgis | erant |
| :--- | :---: | :---: | :---: |
| who:NOM.M.PL | neighbour:NOM.M.PL | Belgian:DAT.M.PL be:IND.IMPF.3PL to |  |
| ea | quae | apud eos | gerantur |
| those things:ACC.N.PL which:NOM.N.PL about they:ACC.M.PL manage:SBJV.PRS.3PL.PASS |  |  |  |
| cognoscant | se | -que de | his |

know:SBJV.PRS.3PL oneself:ACC.SG and about this:ABL.F.PL thing:ABL.F.PL
certiorem faciant
inform:SBJV.PRS.3PL
'He charged the Senones and the other Gauls who were neighbours of the Belgae to find out what was going on among the latter and to inform him thereof'
(7) Tum vero dubitandum non existimavit quin ad

Then actually doubt:GERUNDIVE.ACC.N.SG NEG think:IND.PF.3SG that:CONJ to:PREP eos proficisceretur
they:ACC.M.PL move:SBJV.IMPF.3SG
'Then accordingly he thought that he must no longer hesitate about moving towards them'
8) Re frumentaria provisa
castra movet

Provisions:ABL.F.SG provide:PTCP.PF.ABL.F.SG camp:ACC.N.PL move:IND.PRS.3SG
diebus -que circiter XV ad fines Belgarum pervenit day:ABL.M.PL and about:ADV fifteen to border:ACC.M.PL Belgian:GEN.M.PL arrive:IND.PF.3SG 'After providing his provisions, he moved his camp, and in about fifteen days reached the borders of the Belgae'
Eo cum de improviso celerius $\quad$-que omni opinione
there when suddenly rapidly:COMP and all:ABL.F.SG opinion:ABL.F.SG


Figure 4.6: The dependency tree of sentence (6) from the LDT.



Figure 4.7: The dependency tree of sentence (7) from the LDT.


Figure 4.8: The dependency tree of sentence (8) from the LDT.

Figure 4.9: The dependency tree of sentence (9) from the LDT.

Predicates The predicate(s) of the main clause(s) is (are) given the tag 'Pred'. The predicates occurring in subordinate clauses receive their tags according to the role of the corresponding clause in the sentence. For instance, the predicate of a subject clause is tagged as 'Sb' (subject): node proficiscentur-9 in figure 4.7 is the predicate of the subject clause ad eos proficisceretur, introduced by the conjunction quin-6 and depending on the passive periphrastic dubitandum-3 (esse)..$^{12}$ Accordingly, the predicate of an adverbial clause is tagged as 'Adv': the node inciperet-8 in figure 4.5 is the predicate of the adverbial clause cum primum pabuli copia esse inciperet introduced by the conjunction cum-3. Adverbial clauses are annotated as common substantive adjuncts: see, for example, the neuter adverbial primum-4 modifying the verb inciperet-8 in figure 4.5.

Verbal arguments: 'Sb', 'Obj', 'Pnom', 'OComp' The function of subject is annotated as 'Sb' and can have several morpho-syntactic realisations:

- nominative NPs: node Caesar-3 in figure 4.4, subject of the verb esset-2;
- ablatives in ablative absolutes: node re-1 in figure 4.8; ${ }^{13}$
- finite verbs which are predicates of relative clauses: node sunt-18 in figure 4.9; ${ }^{14}$
- finite verbs which are predicates of subject clauses: node proficisceretur-9, head of the subject clause quin ad eos proficisceretur in figure 4.7;
- accusative nouns in accusative + infinitive constructions: node quam-29 depending on the verb esse-31 in figure 4.4. ${ }^{15}$

The 'Obj' tag is used to annotate both direct and indirect objects, including those arguments that cannot become subjects in passivization, as well as so-called indirect objects with or without a preposition. The morphosyntactic realisations of direct objects include:

[^20]- accusative NPs: node negotium-2 depending on the predicate dat-1 in figure 4.6;
- infinite verbs occurring in accusative + infinitive constructions: node esse-31, depending on dixeramus-34 in figure 4.4;
- infinitive verbs completing modal verbs like possum or volo: node esse-7, depending on inciperet-8 in figure 4.5;
- gerunds and gerundives.

The morphosyntactic realisations for indirect objects include:

- dative NPs: nodes Senonibus-3 and Gallis-6 depending on the verb dat-1 in figure 4.6;
- genitive or ablative NPs: node eo-1 in figure 4.9;
- PPs, where the preposition receives the 'AuxP' tag and the NP receives the 'Obj' tag: nodes $a d-10$ exercitum-11 depending on the verb venit-12 in figure 4.5 .
'Pnom' annotates predicate nominals, namely the determining complements of subjects. They usually occur as nominative NPs. See for example node certior-24 in figure 4.4, referring to Caesar-3 and depending on fiebat-25.
'OComp' is used for annotating object complements, i.e. the determining complements of objects. Object complements usually appear as accusative NPs. See node certiorem-23 referring to se-19 and depending on faciant-24 in figure 4.6.

Prepositions and conjunctions: 'AuxP' and 'AuxC' Following the PDT, in the LDT and in the IT-TB prepositions and conjunctions are viewed as "bridge" auxiliary structures and are annotated respectively with 'AuxP' and 'AuxC'. This implies that, for example, in sentence (10) the pronoun te depends on the preposition ad through the 'Obj' function, that is the dependency relation that would syntactically link te to the verb venio, as illustrated in figure 4.10.
Venio ad te.
go:IND.PRS.1sG towards you-ACC.sG
'I come towards you.'


Figure 4.10: Tree of sentence (10).

Similarly, in subordinate clauses (except for relative clauses, see footnote on page 65) the conjunction is the bridge between the head verb and the verb of the subordinate clause: see node uti-11 in figure 4.4 (page 60 ), introducing the subordinate clause supra demonstravimus.

Coordination and apposition: 'Coord' and 'Apos' As with prepositions and conjunctions, coordinating and apposing constructions are annotated using conventional dependency labels which do not reflect a real syntactic dependency. Therefore, in a coordination the coordinating node (i. e. the conjunction) is the head and its dependents are the coordinated elements appended with the suffix '_Co'. For instance, the coordinating enclitic node que-40 links the two verbs coniurare-39 ('Obj_Co') and dare-44 ('Obj_Co') in figure 4.4. Similarly, in an apposing construction the apposition is the head and its dependents are tagged with the suffix '_Ap'. An apposition (just like a coordination) can appear as a comma as in figure 4.9, where the comma (node 26) separates the coordination et-24 (coordinating Iccium-23 and Andebrogium-25) and primos-27: primos depends on the apposing comma via the object relation, that is the relation it bears to the phrase's head miserunt- 30 .

### 4.5.3 Overview of the database queries



Figure 4.11: Sentence (2) (page 57) of the LDT as it is loaded into the database table for the LDT.

We loaded both corpora into two database tables. Figure 4.11 presents a screenshot of the database interface with the entry of sentence (2) on page 57 from the database table for LDT (see figure 4.2). Each row is a word of the sentence and each column is one of its attributes (like lemma, case, etc.).

In order to extract the SC lexicon, we searched the databases through the MySQL queries, briefly outlined here and described in detail in section A.2.

1. Searching for verbal tokens. Since the lexicon's entries correspond to the verbal lemmas, we first looked for all occurrences of verbs in the treebank. See section A.2.1.
2. Searching for verbal arguments: direct head-argument dependencies. The dependency between the verbal head and the argument node can be direct (as in figure 4.3 on page 58), or indirect (as in figure 4.10). In the first case, we looked for verbal arguments in the treebanks through four functional labels specifically marking them: ‘Sb' (subject), 'Obj’ (direct and indirect object), 'Pnom' (predicate nominal), and 'OComp' (object complement). See section A.2.2.
3. Searching for verbal arguments: indirect head-argument dependencies. In the case of indirect dependencies, "bridge" nodes intervene between the verbal headword
and the its arguments. These intermediate nodes can be prepositions, conjunctions, coordinating elements, or apposing elements. Figure 4.4 on page 60 shows an example of a prepositional node, where the argument eum is attached to the verb adferebatur-19 via the preposition ad. Figure 4.6 on page 62 shows two coordinated objects of the verb dat-1: senonibus-3 and Gallis-6. See sections A.2.3 and A.2.4 for details on the queries and for a description of more complex cases.

### 4.6 Subcategorization representation

In sections 4.5.3 and A. 2 we focussed on the syntactic tags used for annotating the verbal arguments in the treebanks, and we illustrated the different possible combinations of arguments and intermediate nodes in the trees. We finally stressed the importance of retrieving all these combinations so that the SC lexicon completely reflects the verbal valency in the treebanks. After extracting every single occurrence of verbal arguments, these argument nodes must be combined together into SC patterns, one for each verbal token.

In this section and in section A. 3 we describe how we chose to formally represent the SC information in the lexicon, considering the features of the Latin language. Sections A. 3 and 4.6.1 describe the different SC frames and SC classes we defined to collect different information in the lexicon, whereas section 4.6 .2 shows the visual aspect of the lexicon. Finally, section A. 4 presents how frequency information is recorded in the lexicon.

### 4.6.1 Subcategorization frames and classes

As described in section 4.3.3.2, SC lexicons developed for modern languages like English encode the valency information into SC patterns that show the sequence of arguments in the order they appear in the sentences. For example, a transitive instance of a verb can be represented with a 'V_NP' frame.

Considering the relatively free word order of the elements in Latin sentences illustrated in section 4.5.1, we decided to define two main structures in the SC lexicon: SC frames (SCFs) and SC classes (SCCs). We use the term SC structures (SCSs) to refer to both types (frames and classes). The SCFs record the linear order of the arguments in the sentence whereas SCCs do not. The linguistic reason for this choice is that these structures respond to different needs and can be used for different analyses. SCFs prove to be useful in studying the properties of Latin linear order, since it may be interesting, for instance, to analyse when the subjects follow the verbs and when they precede them in our corpora. On the other hand, the SCCs reflect the traditional notion of verbal valency as the number and type of arguments required by a verb and can be used in investigating Latin verbal valency from a more general perspective.

For example, in sentence (3) on page 58 the verb habeo is preceded by a subject (praedicatum) and followed by an object (rationem). The SCSs record precisely this type of information. The SCF and the SCC for the verb habeo in figure 4.3 are respectively:

$$
\begin{aligned}
& \text { SCF: } \mathrm{Sb}+\mathrm{V}+\mathrm{Obj} \\
& \text { SCC: Obj, Sb }
\end{aligned}
$$

In the SCF ' V ' indicates the relative position of the verb with respect to its arguments in the sentence. We used the ' + ' sign to show that the order of the elements in SCF
mirrors the actual order of the words in the original sentence. Conversely, SCC is represented as a list of functional labels ordered alphabetically. When the verb appears with no explicit arguments, as esse-7 in sentence (5) (60), its SCF and SCC coincide and are defined by convention as ' V '.

In addition to the syntactic roles of the arguments, the SCSs display some morphological information regarding the case of nouns, adjectives and pronouns and the mood of verbs occurring as arguments. This content is particularly important for an SC lexicon for Latin, because of the fundamental role played by morphology in argument realisation. In some cases, we decided to add a more fine-grained distinction than the one provided by the treebank annotation. In particular, we distinguished the "ablative subjects" occurring in ablative absolutes and the "accusative subjects" occurring in accusative + infinitive constructions from all the others, and for this we used the special labels 'abs' and 'acc-inf', respectively.

In the previous example, the SCF and the SCC for the verb habeo, enriched with this morphological information, are respectively:

$$
\begin{gathered}
S C F_{\text {morph }}: \mathrm{Sb}[\mathrm{nom}]+\mathrm{V}+\mathrm{Obj}[\mathrm{acc}] \\
S C C_{\text {morph }}: \mathrm{Obj}[\mathrm{acc}], \mathrm{Sb}[\mathrm{nom}]
\end{gathered}
$$

Furthemore, the SCSs contain the lexical fillers of the arguments. The SCF and SCC for habeo enriched with these lexical data in sentence (3) are:

$$
\begin{gathered}
S C F_{\text {lex }}: \operatorname{Sb}\{\text { praedicatum }\}+\mathrm{V}+\operatorname{Obj}\{\text { ratio }\} \\
S C C_{\text {lex }}: \operatorname{Obj}\{\text { ratio }\}, \mathrm{Sb}\{\text { praedicatum }\}
\end{gathered}
$$

This type of data can be used for corpus-based lexical studies and SP acquisition, as we will see in chapters 7 and 5 , respectively.

Finally, the voice ('A' for active, ' P ' for passive) of the verbal token is recorded in the SC lexicon. The distinction between deponent and active verbs in the IT-TB is based on their lemmas only, whereas the LDT distinguishes the actual passive verbal forms from the active ones for deponent verbs as well. Therefore, the IT-TB sublexicon has also a ' $D$ ' (deponent) values for its 'voice' field. This difference is kept in the SC lexicon and needs to be accounted for in all analyses that use the data from these resources. The voice of the verb habeo in sentence (3), along with the syntactic tags, the morphological features and the lexical fillers of its arguments are presented below.

$$
\begin{aligned}
& S C F_{\text {diat-morph-lex }}: \text { A_Sb[nom] }\{\text { praedicatum }\}+\mathrm{V}+\mathrm{Obj}[\mathrm{acc}]\{\text { ratio }\} \\
& S C C_{\text {diat-morph-lex }}: \mathrm{A}-\mathrm{Obj}[\mathrm{acc}]\{\text { ratio }\}, \mathrm{Sb}[\text { nom }]\{\text { praedicatum }\}
\end{aligned}
$$

This discussion has shown the potentialities of the lexicon, where the user can choose the type of information (voice, case, mood, lexical items, etc.) to display. Section A. 3 goes into the details of these choices.

### 4.6.2 The entries of the lexicon

The SC lexicon is built as database tables. Figure 4.12 displays the entry for the verb impono in the SC sublexicon extracted from the LDT, whereas figure 4.13 displays the same information extracted from the IT-TB. Note that the tenth row in figure 4.13 corresponds to sentence (1) (see figure A. 6 on page 178). In section A. 3 we define the different types of SCFs and SCCs. For reasons of clarity and space, we decided to display the $S C C$ $\qquad$ structures only, as they are probably more interesting than the $S C F_{1}$ ones for most readers. Nevertheless, any combination of the $S C C_{\text {lex }}$,

| F. MySQL Query Browser - Connection: rootelocalhost:3306 / Idt_valex2 |
| :--- |
| File Edit View Query |

Figure 4.12: Lexical entry for the verb impono from the SC lexicon for the LDT. For the definition of $S C C_{2}$, see section A.3.
$S C F_{l e x}, S C F_{\text {morph }}, S C C_{\text {morph }}, S C F_{\text {voice }}$, and $S C C_{\text {voice }}$ structures can be displayed in the lexicon, according to the specific needs.

Each entry of these sublexicons is a row of the database table and corresponds to a verbal occurrence (identified by the 'ID' field) in one of the treebanks. In addition, the tables representing the two sublexicons can be merged into one single table, with an additional field called 'treebank' specifying 'IT-TB' or 'LDT' depending on the sublexicon chosen.

Section A. 4 explains how we included frequency information in the lexicon; this allows us to extract quantitative data such as the frequency of a particular SC structure with a particular verb, or the total frequency of a structure, etc. An example of use of such data is given in section 4.7.

### 4.7 Quantitative data

In this section we will present some quantitative data in order to give an overview of the lexicon. We will globally compare the two sublexicons, although we are aware of the fact that their different composition (one author, time period and genre for the IT-TB and several ones for the LDT) may affect the distribution of the observed variables. A statistical study based on these data will give us the chance to investigate this aspect more in detail in chapter 7, showing how the authors in the corpora contribute in different ways to the linguistic analysis on Latin preverbs.

| 7 MySQL Query Browser - Connection: rootelocalhost:3306 / it_valex2 |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: |
| File Edit View Query Script Tools Window MySQL Enterprise |  |  |  |  |
| select $*$ from subcat_lexicon_it where verb="impono" |  |  |  |  |
| $\bigcirc$ Resultset 1 |  |  |  |  |
| author | ID |  | scc2_voice_morph | sentence |
| - Thomas | 976 | impono | P_(a)0bi[abl],Sb[nom] | 001.1SN.DS-4QU2.AR2-RA-4.4-4.6-6 |
| Thomas | 1009 | impono | P_(a)Obi[abl),Sb[nom] | 001.1SN.DS-4QU2.AR2-RA-4.9-1.10-74 |
| Thomas | 4522 | impono | P_(a)Obi[abl),Sb[nom] | 001.1SN. DS22QU1.AR1-RA-3.13-8.18-2 |
| Thomas | 4579 | impono | P_(ad)Obi[acc],Sb[nom] | 001.1SN.DS22QU1.AR2-C0--49-2.52-29 |
| Thomas | 4842 | impono | P_[a)0bi[abl] | 001.1SN.DS23QU1.AR4-RA-4.1-1.3-4 |
| Thomas | 4849 | impono | P_(a)Obi[abl),Sb[nom] | 001.1SN.DS23QU1.AR4-RA-4.3-5.5-84 |
| Thomas | 4866 | impono | P_(a)Obi[abl),Sb[nom] | 001.1SN.DS23QU1.AR4-RA-4.3-5.5-8 |
| Thomas | 4871 | impono | P_(a)Obi[abl),Sb[nom] | 001.1SN.DS23QU1.AR4-RA-4.6-1.7-6 |
| Thomas | 5131 | impono | P_(a)0bi[abl] | 001.1SN.DS25QU1.AR4-C0-.13-8.17-6\% |
| Thomas | 5304 | impono | P_(a)Obi[abl),Sb[nom] | 001.15N.DS25QU1.AR4-RA-1.6-1.7.74 |
| Thomas | 5927 | impono | P_(a)Obi[abl),Sb[nom] | 001.1SN.DS29QU1.AR4-RA-2.3.6.8-6\% |
| Thomas | 26496 | impono | P_(a)Obi[abl),Sb[nom] | 003.3SN.DS-6QU1.AR3-C0--5-1.6-74 |
| Thomas | 26514 | impono | P_Obi[dat], Sb[nom] | 003.3SN.DS-6QU1.AR3-C0--16-4.19-3¢ |
| Thomas | 26520 | impono | P_(a)0bi[abl] | 003.3SN.DS-6QU1.AR3-C0--16-4.19-3 ${ }^{\text {a }}$ |
| Thomas | 37691 | impono | A_Obilacc] | 004.4SN.DS-7QU1.AR3AAG-1.3-6.474 |
| Thomas | 37720 | impono | A_Obi[acc], $\mathrm{Sb}[\mathrm{acc}$ ] | 004.4SN.DS.7QU1.AR3AAG-2.5-1.8-6 |
| Thomas | 44733 | impono | P_Obidat] | 004.4SN.DS17QU1.AR2CCO--.14-4.22-7\% |

Figure 4.13: Lexical entry for the verb impono from the SC lexicon for the IT-TB.


Figure 4.16: Distribution of verb frequencies in the two sublexicons.

Section 4.7.1 analyses the distribution of verb frequencies in both sublexicons, while section 4.7.2 focusses on the number of different SCSs defined. Finally, section 4.7.3 reports on a small study on word order patterns extracted from the lexicon. See section A.4.2 for an analysis on the frequency of co-occurrence between verbs and SC structures.

### 4.7.1 Verb frequencies

The six most frequent verbs occurring in the treebanks are listed in table 4.7. Table 4.6 shows the number of different verbal tokens (forms and lemmas) $|T|$ and the number of verbal types (forms $\left|V_{T f}\right|$ and lemmas $\left|V_{T l}\right|$ ) in the LDT and in the IT-TB lexicons. From tables 4.7 and 4.6 we notice that - perhaps not surprisingly - the most frequent

| treebank | author | $\|T\|$ | $\left\|V_{T f}\right\|$ | $\frac{\left\|V_{T f}\right\|}{\|T\|}$ | $\left\|V_{T l}\right\|$ | $\frac{\left\|V_{T T}\right\|}{\|T\|}$ |
| :--- | :--- | ---: | ---: | ---: | ---: | ---: |
|  | Caesar | 295 | 247 | 0.84 | 169 | 0.57 |
|  | Cicero | 1283 | 933 | 0.73 | 458 | 0.36 |
|  | Jerome | 1770 | 774 | 0.44 | 274 | 0.16 |
| LDT | Ovid | 1092 | 845 | 0.77 | 474 | 0.43 |
|  | Petronius | 2901 | 1748 | 0.60 | 731 | 0.25 |
|  | Propertius | 1039 | 767 | 0.74 | 400 | 0.39 |
|  | Sallust | 2294 | 1509 | 0.66 | 631 | 0.28 |
|  | Vergil | 511 | 460 | 0.90 | 317 | 0.62 |
|  | TOTAL | 11186 | 5916 | 0.53 | 1631 | 0.15 |
| IT | Thomas | 9072 | 2108 | 0.23 | 560 | 0.06 |

Table 4.6: Number of verbal tokens ( $|T|$ ) and verbal types (forms $\left|V_{T f}\right|$ and lemmas $\left.\left|V_{T i}\right|\right)$ in the LDT and in the IT-TB sublexicons. $\frac{\left|V_{T f}\right|}{|T|}$ and $\frac{\left|V_{T l}\right|}{|T|}$ show the type/token ratios.
verbs are the same for both treebanks, but Saint Thomas uses much fewer verb types than the authors in the LDT (see the row headed 'TOTAL' in table 4.7), which means that the lexical richness of verbs is lower in the IT-TB than in the LDT. This result is confirmed when we compare the type/token ratios given by the following formulae and displayed in the last column of table 4.6:

$$
\begin{equation*}
\frac{\left|V_{T f}\right|}{|T|} \quad \frac{\left|V_{T l}\right|}{|T|} \tag{4.1}
\end{equation*}
$$

These ratios are much lower for the IT-TB compared to the LDT and this difference must be taken into account when analysing the quantitative data for the two sublexicons. ${ }^{16}$

The distribution of the verb frequencies has a long tail of low-frequency verbs in both corpora (see figure 4.14 for the LDT and 4.15 for the IT-TB). We used the $z i p f R$ package (Evert and Baroni, 2006) to model these frequency distributions and found that the Zipf distribution is not a good model for neither of them. Instead, the Generalized Inverse Gauss-Poisson (GIGP) Large Number of Rare Events (LNRE) model is a fair model for the frequencies of the verbs in the IT-TB $\left(\chi_{(10}^{2}=20.21372, p\right.$-value $\left.=0.027\right)$.

| treebank | verb | frequency |
| :--- | :--- | :--- |
| LDT | sum | 1,535 |
|  | habeo | 308 |
|  | dico | 268 |
|  | facio | 261 |
|  | video | 247 |
|  | possum | 179 |
| IT-TB | TOTAL | 2,798 |
|  | sum | 2,475 |
|  | dico | 795 |
|  | habeo | 372 |
|  | possum | 318 |
|  | facio | 205 |
|  | video | 152 |
|  | TOTAL | 4,317 |

Table 4.7: Six most frequent verbs in the LDT and in the IT-TB. The lemma dico refers to the verb dico, -is 'say' (third conjugation class) and not to dico, -as 'dedicate' (first conjugation class)

The lexicon records the voice of each verb token, and identify the zero-valency usages. Table 4.8 gives the number of lexicon entries (that is the number of verbal tokens), the number of "empty" structures (i.e. the number of times a ' V ' structure is found), and the number of active and passive verbal forms, ${ }^{17}$ with the corresponding percentages

[^21]over all the tokens. Table 4.8 shows that a fourth of the total verbal occurrences consists of these empty frames and that most verbs occur in the active voice.

| treebank | verbal tokens | empty frames |  | active |  | passive |  |
| :--- | :---: | :--- | :---: | :---: | :---: | :---: | :---: |
| LDT | 11186 | 2513 | $22 \%$ | 8969 | $80 \%$ | 2212 | $20 \%$ |
| IT-TB | 9072 | 2235 | $25 \%$ | 6244 | $69 \%$ | 2828 | $31 \%$ |

Table 4.8: Counts of verbal tokens (lemmas), zero-valency usages, active and passive verbal forms in the LDT and in the IT-TB sublexicons.

### 4.7.2 Subcategorization frequencies

In section 4.6 we described the two main types of SCSs making up the lexicon's entries: SCFs (frames) and SCCs (classes). In order to show the different granularity of these structures, and hence their different distribution, in table 4.9 we reported the total number of the different structures. ${ }^{18}$ For each structure, this count is higher in the LDT than in the IT-TB lexicon, when compared to the size of the treebanks: the results of a Pearson $\chi^{2}$ tests of indipendence are indeed significant $\left(\chi_{(1)}^{2}=78.84\right.$ for SCF, $\chi_{(1)}^{2}=23.40$ for SCC, $2.2 e-16<p<1.607 e-05$ in both cases), ${ }^{19}$ but the effect sizes are very small $(0.008<\varphi<0.027)$, leading us to attribute the observed differences to expected variation in sampling.

|  | $S C F$ | $S C C$ |
| :--- | :---: | :---: |
| LDT | 655 | 238 |
| IT-TB | 384 | 148 |

Table 4.9: Counts of the different SCSs in the LDT and in the IT-TB sublexicons. More precisely, the $S C F$ and $S C C$ structures referred to here are $S C F_{1_{\text {voice }}}$ and $S C C_{2_{\text {voice }}}$, respectively; for the definition of $S C F_{1}$ and $S C C_{2}$, see section A.3.

An overview of the 15 most frequent SCC structures recorded in the sublexicons is displayed in table 4.10, whereas table A. 10 on page 186 shows the 5 most frequent SCC structures by author.

From table 4.10 we can see an expected large number of what we called "empty" structures ('A_V' or 'P_V' depending on the voice; see also table 4.8). In these cases the verb occurs with no expressed arguments. For example, the sentence Dixerat, et flebant (Ovid, Met., I, 367) "He spoke, and they were crying" displays a zero-valency usage of the verb 'dico' in the sense of 'talk', while the valency structure of dico in the usual sense of 'say' would require (a subject and) an object.

Furthermore, SCSs containing prepositions do not appear among the most frequent SCSs, which only include various combinations of ' Sb ' and ' Obj ' arguments, with 'Pnom' occurring quite often in the IT-TB lexicon.

[^22]| LDT | IT-TB |  |  |
| ---: | :--- | ---: | :--- |
| A_V | 1669 | P_V | 1188 |
| A_Obj[acc] | 1668 | A_V | 1047 |
| A_Obj[acc],Sb[nom] | 1052 | A_Pnom[nom],Sb[nom] | 955 |
| A_Sb[nom] | 963 | A_Sb[nom] | 855 |
| P_V | 843 | P_Sb[nom] | 629 |
| A_Obj[inf] | 363 | A_Obj[acc],Sb[nom] | 578 |
| P_Sb[nom] | 294 | A_Obj[acc] | 499 |
| A_Pnom[nom],Sb[nom] | 274 | A_Pnom[nom] | 346 |
| A_Obj[inf],Sb[nom] | 260 | A_Obj[inf],Sb[nom] | 315 |
| P_Sb[abl] | 171 | A_Obj[inf] | 148 |
| A_Pnom[nom] | 156 | P_Pnom[nom],Sb[nom] | 122 |
| A_Obj[acc],Obj[dat] | 147 | P_Obj[dat],Sb[nom] | 101 |
| P_Obj[abl] | 126 | A_(quod)Sb[sbjv] | 86 |
| A_Sb[acc] | 112 | P_Obj[dat] | 79 |
| A_Obj[dat],Sb[nom] | 105 | P_(a)Obj[abl] | 73 |

Table 4.10: 15 most frequent types of $S C C_{\text {voice-morph }}$ and their frequencies in the LDT and IT-TB lexicons. More precisely, the SCC structures referred to here are $S C C_{2_{\text {voice-morph }}}$; for the definition of $S C C_{2}$, see section A.3.

As to the morphological information (case and mood) for these functional labels, we find that verbal arguments are morphologically expressed as expected in most cases: nominative for subjects and predicate nominals, accusative for direct objects and object complements, dative and ablative for indirect objects. However, table A. 11 on page 187 indicates the important role played by alternative constructions, such as ablative absolutes ('Sb[abl]'), accusative + infinitive clauses ('Sb[acc]' and 'Obj[inf]') and subject clauses ('Sb[ind]' and 'Sb[sbjv]'). In addition, subjects in genitive correspond to gerundives, while nominative objects in most cases belong into reported speech or verb forms in participle.

In order to improve the quality of the treebanks and therefore the quality of the input for the subsequent studies, we manually checked the "unespected" cases, such as objects in nominative or vocative subjects. This allowed us to correct some inconstistencies in the treebank annotation and in the morphological analysis, which were particularly numerous in the LDT.

The functional labels of the most frequent intermediate nodes betwen the verb and its arguments are displayed in table 4.11. We note that coordinations and prepositions are the most frequent intermediate nodes between verbs and their arguments, followed by appositions and conjunctions. The most frequent lemmas is et in both lexicons; then come other coodinating elements (que and atque) ${ }^{20}$ in the LDT and subordinating conjunctions (quod) and prepositions ( $a, a d, i n$ ) in the IT-TB.

### 4.7.3 Word order patterns

In order to show the use of the SC structures we defined in section 4.6.1, in this section we report on a small study on word order patterns extracted from the lexicon.

[^23]| LDT |  |  |  | IT-TB |  |  |  |
| ---: | :--- | ---: | :--- | ---: | ---: | ---: | ---: |
| func.label | freq | lemma | freq | func.label | freq | lemma | freq |
| Coord | 1781 | et | 824 | Coord | 836 | et | 494 |
| AuxP | 759 | que | 417 | AuxP | 790 | quod | 332 |
| Apos | 154 | in | 231 | AuxC | 374 | a | 268 |
| AuxC | 67 | atque | 200 | AuxC,Coord | 79 | ad | 237 |
| AuxP,Coord | 62 | ad | 147 | Apos | 59 | in | 178 |

Table 4.11: Most frequent functional labels (syntactic functions) and lemmas of the intermediate nodes between the verbs and their arguments.

| pattern | LDT lexicon | IT-TB lexicon |
| :--- | ---: | ---: |
| OV | 153 | 37 |
| VO | 116 | 40 |

Table 4.12: Frequencies of the word order patterns ' VO ' and ' OV ' in the two sublexicons.

In section 4.6.1 we defined SC frames (SCFs) as SC structures that record the order of the elements. An example of an SCF is ' $\mathrm{A} \_\mathrm{Obj}+\mathrm{Obj}+\mathrm{Sb} \_\mathrm{Co}+\mathrm{Sb} \_\mathrm{Co}+\mathrm{Sb} \_\mathrm{Co}+\mathrm{V}$ ' which means that we have two objects followed by three coordinated subjects and an active form of a verb.

For this study we extracted all SCFs from our lexicon according to the following criteria: the voice of the verb form was active, the SCF only contained the labels ' V ' for verbs (obligatory), 'Sb' (facultative), and 'Obj' (obligatory). In addition, we excluded the cases where two or more objects occured to the left and to the right of the verb at the same time. We collected the frequencies of these frames in each treebank.

We did not want the data to be biased by the fact that Latin is a pro-drop language; consequently, we focussed on the alternation between the patterns ' VO ' and ' OV '. To do this, we produced a script to automatically associate all frames to one of the two patterns. For example, ' $\mathrm{Obj}+\mathrm{Obj}+\mathrm{Sb} \_\mathrm{Co}+\mathrm{Sb} \_\mathrm{Co}+\mathrm{Sb} \_\mathrm{Co}+\mathrm{V}$ ' is reduced to 'OV'.

Table 4.12 shows the frequencies of the two patterns in the two sublexicons. In order to see if there is a significant association between the patterns and the sublexicons, we performed a $\chi^{2}$ test of independence: see section C. 2 for an illustration of this test. The result was not statistically significant $\left(p=0.2141, \chi_{(1)}^{2}=1.54\right)$. This might have been due to the fact that the data for the LDT sublexicon refer to a number of authors belonging to very different genres and ages. This is confirmed by the fact that if we merge the data for Jerome (the latest author in the LDT) with those for Thomas (see table 4.13), we get a significant result ( $p<0.01, \chi_{(1)}^{2}=77.79$ ); we also get a large effect ( $\varphi=0.474$ ), which means a strong association between the word order patterns and our choice of classification by author groups.

The result of this case study shows that the data associated with Jerome affect the outcome of the test: the presence of Jerome in the first test pulls the distribution of the LDT data closer to the IT-TB data, which might indicate a diachronic trend.

In chapter 7 we will make use of more advanced statistical techniques that are able to account for the contribution of different variables - such as the author, the specific verb, the type of argument - to carry out a study of Latin preverbs.

| pattern | LDT without Jerome | IT-TB +Jerome |
| :--- | ---: | ---: |
| OV | 152 | 38 |
| VO | 52 | 107 |

Table 4.13: Frequencies of word order patterns in the LDT lexicon without the data for Jerome, contrasted with the IT-TB sublexicon plus the data for Jerome.

### 4.8 Remarks and future work

In this section we summarise the main aspects of the SC lexicon presented in this chapter, stressing their value and limits, and presenting future directions of research. In particular, section 4.8 .1 collects remarks on the features of the SC lexicon and contrasts it with other comparable resources. Finally, section 4.8.2 outlines some future developments of this research.

### 4.8.1 Features of the lexicon

The SC lexicon described so far displays information on the argument structure of Latin verbs; these data are extracted on the basis of the actual occurrences of verbs and their arguments in the treebanks we used (LDT and IT-TB).

Our definition of what constitutes an argument is therefore purely usage-based in the sense that it fully relies on the way the corpora were semi-automatically annotated. In our extraction system, verbal arguments are simply defined as those nodes depending on a verb (sometimes via certain intermediate nodes) and annotated with some specific syntactic labels (subject, object, etc.). The distinction between arguments and adjuncts described in section 2.2 is already present in the original annotated data and does not need to be acquired in the extraction step. Consequently, according to the distinction between corpus-based approaches and corpus-driven approaches proposed in Tognini-Bonelli (2001), our lexicon can be seen as a corpus-based resource because it automatically detects the structures contained in the corpora on the basis of their annotation, which relies on a specific theoretical framework (Functional Generative Description, see page 21 ff .); the extracted information is thus affected by the annotation schema followed. On the other hand, previous automatically acquired SC dictionaries such as those described in section 4.3.3 exploit large corpora consisting of several million words in order to achieve higher accuracy while employing statistical techniques and automatic parsers. In that context, the distinction between arguments and adjuncts is defined in the acquisition phase, since the goal is to be able to filter out modifiers and extract verbal arguments only.

Finally, the layout of the lexicon shows how the treebank annotation can in turn be improved. Some choices adopted in the treebank annotation may look questionable to some readers and thus need a further explanation. This difficulty is, of course, reflected in the lexicon, which cannot intervene in the annotation phase itself, but uses the annotation as input. For example, as mentioned earlier (sections 4.5.2.2 and A.3), subjects occurring in ablative + infinitive constructions are syntactically tagged as ' Sb ' and morphologically marked as ablatives. However, their morphological realisation does not depend on their being subjects, but on the adverbial nature of the clause they are embedded in. Even though these cases are easy to detect in the lexicons making its content unambiguous (because they correspond to the instances of ' $\mathrm{Sb}[\mathrm{abl}]^{\prime}$ ), this may still be considered conceptually misleading by some linguistic approaches.

### 4.8.1.1 Comparison with traditional and modern dictionaries

The way we exploited the original corpora while creating the lexicon distinguishes it both from traditional and modern lexical resources of the same kind.

First, the tight connection between the lexicon and the corpora it was obtained from differentiates it from traditional Latin dictionaries and thesauri. The latter ones typically aim at illustrating the behaviour of words by providing specific examples from text in a non-systematic fashion (see section 2.2). At the same time, these resources usually give a broad overview of Latin, without limiting the analysis to particular works or authors. Among these, Happ (1976a, pp. 480-565) generalizes an analysis based on a small corpus to Classical Latin as a whole.

In contrast, our SC lexicon aims at comprehensively describing the linguistic evidence represented by the two treebanks we used. For this reason, the detected SC structures should not be intended as representing the general behaviour of Latin verbs' arguments: they rather give a picture of the distribution of verbal arguments in the particular corpora at hand. When generalizing these results - as described in chapter 5, for instance - the special composition of the treebanks (time periods, authors, genres) should be considered as an important factor.

In addition, the quantitative data on frequency are not found in traditional dictionaries and represent valuable information to be used for further studies, as we will explain later on (section 4.8.2).

Compared with the two previous examples of SC lexicons for Latin (Happ, 1976a; Bamman and Crane, 2008), our lexicon presents several advantages. First, its size. It is extracted from two medium-sized treebanks amounting to close to 100,000 word tokens and contains lexical entries for around 20,000 verb tokens. Happ (1976a)'s lexicon was obtained from a manual inspection of 800 verbal occurrences in Cicero, while Bamman and Crane (2008)'s lexicon was extracted from an automatically parsed corpus of 3.5 million words. The small size of Happ's corpus is somehow balanced by the high quality of the data, which were manually analysed, but it hardly allows for generalization. On the other hand, the noise present in Bamman and Crane's corpus is compensated for by its large size and the use of statistical techniques. We propose a lexicon which is obtained from high quality data and, at the same time, is relatively large: its size was manageable thanks to the automatic methods we used to extract it. In addition, its high quality makes it possible to perform linguistic studies which target low-frequency phenomena, as well as medium and high-frequency phenomena.

### 4.8.1.2 Use of the lexicon

Dealing with an ancient language affects the aims and the audience of the research we are carrying out. If corpora for modern languages are often exploited in computational linguistics as training sets to automatically analyse new unseen texts, corpora for ancient languages combine this aim with the traditional philological interest in the text per se and in language change.

In this perspective, the SC lexicon we extracted can contribute both to Latin corpus linguistic research and to research in NLP for Latin. For what concerns the first, in chapter 7 we will present a linguistic analysis based on the lexicons; regarding the second, in chapter 5 we will proceed towards a further (semantic) processing of the Latin language, starting from the information contained in the lexicon.

The automatic procedure The system for automatic extraction of SC described in this chapter also has some important practical advantages. It makes it possible to update the lexicons as the treebanks increase just by running the database queries again. Furthermore, the procedure can also be extended to other treebanks that share the same annotation guidelines, such as the PROIEL Project Treebank (Haug and Jøndal, 2008) and the recently initiated CoLaMer project. ${ }^{21}$ Finally, it gives a way to check the consistency of the annotation itself, as we explained earlier (see section 4.7.2).

### 4.8.2 Future directions of research

The lexicon can still be improved in many ways and these improvements can enhance its usefulness in several applications.

First, the corpus-based procedure does not allows us to collect single optional syntactic slots into the SC patterns, thus recording the so-called phenomenon of frame alternation. For example, from table A. 9 on page 185 we can infer that the verb do can appear with a combination of an optional ablative object, a dative object and a nominative subject. This could be represented as the following SCFs: "Obj[abl] OPT, Obj[dat] OPT, Sb[nom]OPT", where we tagged the optional slots as 'OPT'. Such a representation could be produced in an automatic way, while at the same time improving the clarity of the lexicon for human readers.

Second, as we have already said, the lexicon fully relies on previously annotated texts, and thus its accuracy depends on the quality of these data. The next step would consist on a formal evaluation of the lexicons. For this we could compare them with other existing similar resources for Latin, such as Happ's list, the Perseus "dynamic lexicon", traditional dictionaries and thesauri (chapter 2).

Third, since the treebanks already annotate arguments of adjectives, the lexicon could be enriched with valency information for adjectives by using queries similar to the ones we have already designed.

Moreover, since dictionaries often record the adjuncts most commonly used with a verb, besides its arguments (see chapter 2), we may add these elements in the lexicon, after setting an appropriate frequency threshold.

Also, the lexicon would gain even more value if it associated each lexical entry with the verb sense rather than the verb form. This is now partially achieved because the LDT indicates the different verb senses based on the word lemmas of the Lewis and Short (1879) dictionary. For what concerns the IT-TB, adding this information would require a Word Sense Disambiguation system.

The lexicon be employed in improving the performance of statistical parsers for Latin, ${ }^{22}$ as the valency information it displays may be included among the features specific to Latin syntax for syntactic parsers (see Carroll et al. 1998; Hlaváčková and Kadlec 2006). Moving from parser performance to lexicon development, the lexicon can be extracted from automatically parsed texts when an accurate parsing system is available.

### 4.9 Summary

In this chapter we presented the Latin SC lexicon we have built. We explained the motivation that led to this work in relation to the state of the art in (computational)

[^24]lexicography. We illustrated the DG annotation style characterizing the two Latin treebanks used and explained the procedure for the extraction of the lexicon. Moreover, we described the formal and visual aspect of this resource and gave a quantitative overview of the distribution of verbal arguments therein.

Then, we outlined the main features that relate the lexicon with other comparable lexical resources developed for modern languages and for Latin. Finally, we showed how it may be improved and used in corpus linguistics and NLP applications. An example of use in the first is given in chapter 7, whereas in chapter 5 we deal with one of the possible directions in the latter.

## Chapter 5

## Selectional Preferences

### 5.1 Introduction

This chapter describes an unsupervised probabilistic model for automatically acquiring selectional preferences (SPs) from syntactically annotated Latin corpora using Latin WordNet (LWN) as the external reference ontology (Minozzi, 2009). This work, which relies on the theoretical definitions and motivations given in chapter 1 and follows the approach by Alishahi (2008), complements the research reported in chapter 4 by integrating a lexical semantic module into the syntactic module already included in the subcategorization (SC) lexicon.

SPs were introduced in section 2.2 .1 on page 15 and can be briefly defined as the semantic preferences of verbs on their arguments. For example, the subject position of the verb $f l y$ is usually filled by lexical items whose semantic properties include being a certain kind of bird or a certain kind of vehicle.

As we will see, previous models for SP acquisition calculate this prefererence by using single verb-argument occurrences collected from large corpora. Since our dataset is extracted from two relatively small treebanks - amouting to around 100,000 tokens overall - a relatively high number of low frequency verb-noun instances are observed. Nevertheless, these low frequencies stem from a small sample which underestimates the variance, and they are not necessarily a property of Latin verbs. Consequently, the generalization step from these verb occurrences to broader classes of nouns may be compromised by the small size.

In order to (partially) remedy this difficulty we can group the observations into larger entities: constructions. Our definition of 'construction' follows from Alishahi (2008), who builds on the concept of form-meaning pair as defined in Construction Grammar (Goldberg, 1995) and extends it to "a grouping of verb usages which share similar syntactic and semantic features, inducing a probability distribution over the values for these features" (Alishahi, 2008, p. 13). We tested to what extent the model is able to generalize observed patterns of association to unseen events both by using these constructions and by extending the set of observations from which the generalization is drawn.

The rest of the chapter is organized as follows. Section 5.2 reports on previous work on SP acquisition. Section 5.3 describes our system for SP acquisition. Section 5.4 reports and discusses our results, while section 5.5 sketches future work and concludes.

### 5.2 Previous work

Over the last two decades research into automatic acquisition of SPs from corpora has led to a number of different systems. Automatic methods proved to be faster and more reliable than manual ones, since they allow to process large amounts of data and are not prone to the inconsistencies typical of manual work. In addition, automatic methods are particularly suited in the case of computationally less-resourced languages such as Latin. For these reasons, in the present section we will focus on automatic methods in SP acquisition and will give a survey of the main contributions in this field. For a more in-depth overview of the literature, we refer to McCarthy (2001) and Light and Greiff (2002), whereas further details about the individual methods mentioned in this section are included in section B.3.

As we saw in chapter 1, automatic acquisition of SPs from corpora focusses on a predicate $v$ and an argument position $r$. Then, the common procedure consists of two main phases:

1. extracting argument headwords from a corpus;
2. generalizing corpus frequencies over unseen cases.

The first phase requires that some kind of information on argument structure is contained in the corpus being used; this information can be extracted from an SC lexicon like the one described in chapter 4 . This way, all noun fillers $n$ occurring in the $r$ relation with $v$ can be collected; they are typically included in triples

$$
(v, r, n)
$$

For example, we could extract the subjects of $f l y$ in our corpus - bee, bird, crow and eagle - and represent them as triples: (fly, subject, bee), (fly, subject, bird), (fly, subject, crow), (fly, subject, eagle).

The second phase is required by class-based methods, since it uses the argument fillers previously extracted to extend the knowledge acquired from the observed cases to unseen ones; this way, one or more noun classes are obtained that represent the verb's argument position at hand. ${ }^{1}$ For example, we could represent the subject position of fly with the classes BIRD, INSECT or even ANIMAL, thus predicting (with different degrees of accuracy) the likelihood of finding (words belonging into) these classes as subjects of fly.

Class-based methods fall into two main groups: knowledge-based approaches and distributional approaches, depending on whether the generalizing classes are taken from an existing taxonomy such as WN, or are automatically extracted from the corpus, without relying on external knowledge bases. In sections 5.2.1 and 5.2.2 respectively we will discuss these two approaches.

### 5.2.1 Knowledge-based approaches

In section 5.2 we saw that knowledge-based approaches rely on ontologies such as the WN noun hierarchy to find the appropriate noun classes for each verb-argument

[^25]pair during the generalization phase: consequently, SPs are represented through noun classes corresponding to WN nodes.

The practical advantages of using WN are well known and concern its free availability and coverage (see section 3.3.1). Furthermore, the fact that WN is structured in sense sets makes it particularly suited for semantic tasks that are tightly connected with SPs, such as Word Sense Disambiguation (WSD).

The main disadvantages of relying on WN for SP acquisition are its non balanced granularity and its manual compilation, which brings with it all the problems connected with human errors.

Next, we will describe the common procedure adopted by the WN-methods for SP acquisition (sections 5.2.1.1 and 5.2.1.2), before giving a more detailed account of them in section 5.2.1.3.

### 5.2.1.1 Overview of the methods

The input data to a typical SP acquisition system are collections of verb-noun cooccurrences in specific syntactic contexts and the aim is to extract broader classes from these nouns. The usual procedure moves upwards from the hierarchy leaves (associated with the fillers) to parent nodes (classes), the main issue being to decide at which level we should stop this process.

For example, let us suppose that the verb $f l y$ appears in our corpus with the following fillers for the subject position: bee, bird, crow and eagle. From now on, we will hypothesize that these four nouns are represented as leaves in the WN-like tree of figure 5.1. ${ }^{2}$


Figure 5.1: Example portion of a WN-like hierarchy for some possible subjects of $f l y$.

The node BIRD $=\{\operatorname{bird}\}^{3}$ is the parent node of bird, crow, eagle and swallow, while INSECT $=\{$ insect $\}$ is the parent node of bee, bug and insect. Together, BIRD and insect could well represent the class of fillers for the subjects of $f l y$ in our corpus, thus generalizing over a large number of unseen events identified by other names of insects and birds. At a higher level, animal is the parent node of all the fillers in our corpus and constitutes a broader class than (BIRD, INSECT): although it is more general, ANIMAL also includes beings that cannot fly, which are not represented in figure 5.1. At the top level - not represented in figure 5.1 - we could hypothesize an even more general class

[^26]such as Entity, which would cover all possible subjects of $f l y$, as well as the arguments numerous other verbs: consequently, the class ENTITY would probably be too broad for our purpose, since it would not be able to differentiate the arguments of the verb we are interested in from all the other arguments.

Therefore, a good generalization should at the same time be comprehensive and reflect the features of the fillers it generalizes from.

In general, the decision on the level of generalization to choose can be taken by looking at the mapping

$$
\text { pref }:(v, r, c) \mapsto p
$$

where the weight $p$ is a real number that corresponds to the degree of preference of $v$ for the class $c$ in relation $r$ and is given by the SP acquisition system. As we will see, in a probabilistic framework these weights appear as probabilities.

### 5.2.1.2 Common procedure

Since the starting point consists of observations in a corpus, corpus frequencies are used for determining the preference weights, by stressing the importance of highly frequent co-occurrences over less frequent ones; at the same time, this reduces noise in the training data. We can thus say that knowledge-based approaches follow two main substeps in the generalization phase:
(a) populating the ontology with corpus frequencies;
(b) estimating the preference weights of ontology classes.

In the ontology-population step, the lexical fillers of the arguments collected from the corpora are associated with leaves in the hierarchy, while their parent nodes correspond to noun classes. This step adapts corpus frequencies of noun lemmas to the particular structure of the ontology chosen (usually WN), which consists of synset nodes linked together though the IS-A relation (cfr. section 3.3.1). One of the main issues here involves noun ambiguity, i.e. the possibility that a noun belongs into more than one node: since its sense is not annotated in the input text, we need to decide how to distribute its frequency among its different senses within the hierarchy. Additionally, it is not obvious how the frequencies should be distributed among the noun's hypernyms: for example, which fraction of the frequency of woman as subject of think should its hypernym PERSON receive?

In the second step, corpus data are used to decide on the preference weights. Thanks to specific statistical tools, the preferred arguments for a verb can be represented as a probability distribution over WN noun classes: this way, different probabilities would be assigned to BIRD and PERSON as subjects of $f l y$, rather than identify BIRD as the only class for $f l y$ 's arguments.

In section 5.2.1.3 we will briefly see how the different systems choose to perform these tasks.

### 5.2.1.3 Survey of the systems

In this section we will report on some WN-based systems for SP acquisition from corpora.

Resnik's model (Resnik, 1993, 1997) - later improved by Ribas (Ribas, 1995) - is the first class-based method for inducing SPs from corpora co-occurrences using an external wide-coverage semantic taxonomy like WN. His system produces selectional association
scores to be associated to each WN node. Following Resnik's approach, other works (Li and Abe, 1998; Clark and Weir, 1999; Abney and Light, 1999; Ciaramita and Johnson, 2000; McCarthy, 2001; Kruengkrai et al., 2004) developed specific statistical models for SPs that deal with noun ambiguity in different ways. These developments will be presented later in this section. Afterwards, we will briefly touch upon an interesting model for class-to-class SP acquisition (Agirre and Martinez, 2001) and finally present three models for SPs developed for specific tasks (Schulte im Walde, 2004; Calvo and Gelbukh, 2004; Pantel et al., 2007).

Resnik (1993) In Resnik's model the WN-population step is performed by uniformly associating an argument filler to all the classes it belongs into, counting from the leaf nodes up to the top nodes. Then, the frequency of a class $c$ is given by summing the frequencies of all nouns $n$ contained in $c$ or in nodes under $c$ divided by the number of classes to which $n$ (directly and indirectly) belongs. In the example of figure 5.1 the frequency of the class BIRD $=\{$ bird $\}$ would be given by summing the contribution of the nouns in it and under it: bird, crow, eagle and swallow; for instance, the contribution of crow would be obtained by dividing the frequency of crow by the number (3) of classes to which crow directly and indirectly belongs: $\{c r o w\}, \operatorname{BIRD}=\{$ bird $\}$ and ANIMAL $=\{$ animal $\}$.

After populating WN with corpus frequencies from the Penn treebank (Marcus et al., 1993), Resnik applies the measure of relative entropy (Kullback and Leibler, 1951) from Information Theory to represent the strength of a predicate's selection for an argument as the quantity of information carried by the verb about that argument (association score). This measure contrasts the "co-occurrence behaviour" of a given verb and a WN class with that of the class alone. More specifically, given an argument position $r$ and a WN class $c$, the prior probability $P_{r}(c)$ refers to the probability of $c$ occurring in the position $r$ with any predicate; following the example given in Resnik (1997), the class PERSON has a much higher probability to occur in a subject position than insect. On the other hand, the posterior probability $P_{r}(c \mid v)$ of $c$ given a verb $v$ relates $c$ (and $r$ ) with $v ;^{4}$ for example, the posterior probability of INSECT with the verb buzz is higher than the same probability for PERSON. ${ }^{5}$ It is precisely this contrast between prior and posterior probability that defines SPs in Resnik's approach: if the verb and the class are strongly associated in relation to the given argument position, then the class will appear more frequently with the verb that it does in general. Resnik's model is described more in detail in section B.3.1.

Ribas (1995) Ribas' proposal aims at solving some of the weaknesses in Resnik's model, mainly derived from sense ambiguity and data sparseness.

As we saw, in Resnik's system the frequency of a noun $n$ is uniformly distributed among all the classes $n$ belongs into, regardless of the difference between direct membership (homonyms of $n$ ) and indirect membership (i.e. parent nodes or hypernyms of $n$ ). On the other hand, Ribas proposes to divide the frequency of a noun by the number of classes into which the noun belongs directly, thus differentiating between synonyms and hypernyms, as illustrated in section B.3.2.

Statistical models Resnik and Ribas present ways to calculate association scores, but do not provide proper statistical models for SPs. Here we will present a group

[^27]of works that produce such models by respectively referring to Information Theory, statistical testing, and Bayesian inference.

Li and Abe $(1995,1998)$ formalise the problem of generalizing from observations to SPs in the context of tree cut models (TCMs), i.e. ways of partitioning a given thesaurus tree, such as the WN hierarchy. ${ }^{6}$ Therefore, their system looks for partitions which disjointly cover all leaf nodes across the hierarchy and estimates the optimal TCM of the given thesaurus tree, for example [BIRD, bug, bee, insect]. For this reason, their definition of SPs differs from that by Resnik and Ribas, who both look for disjoint classes in WN with the highest association score, for example BIRD. The generalization step in Li and Abe's method is based on the Minimum Description Length principle, a principle of data compression from Information Theory (Rissanen, 1989); they thus attempt to find the optimal TCM by choosing the model that most concisely describes the data. The details of Li and Abe's model are given in section B.3.3.

McCarthy (2001) enriches Li and Abe's system with a WSD system and performs a few modifications regarding the frequency contributions of the WN nodes, thresholding, input data from an SC lexicon, and automatic detection of proper nouns. The details of her system are given in section B.3.4.

Clark and Weir (1999) solve the generalization step by means of hypothesis testing on corpus frequencies mapped to WN classes. In their iterative procedure, the first cut along WN contains all the leaf classes; then, the cut is moved up in the hierarchy. The leaf nodes (say, animal-flesh and beverage) are coalesced into the parent node (say, food) if the probability of a food sense occurring in a particular argument position for a verb is independent of the subclass it belongs to: this is decided on the basis of a $\chi^{2}$ test. See section B.3.5 for further details.

In contrast, Abney and Light (1999) introduce a word-based model for SP induction that uses Hidden Markov Models (HMMs) to generate distributions over words and classes on the basis of the co-occurrence probabilities. Each verb-relation $(v, r)$ pair is associated to an HMM shaped as WN, but identified by the specific parameters estimated from the co-occurrence frequencies $f r(v, r, n)$ : the states of the model correspond to the semantic classes in the hierarchy. Starting from the hierarchy root, a transition to a child class is chosen based on the HMMs transition probabilities; this continues until a leaf node is reached and then a noun lemma is generated. Training the model on a $(v, r)$ pair gives a distribution $(p(c \mid v, r))$ over classes and training the model on all nouns gives the distribution $p(c)$. Then, Resnik's definition of association scores can be applied and SPs obtained. See section B.3.6 for a more detailed description.

Finally, Ciaramita and Johnson (2000) use Bayesian networks with the same topology as WN in order to obtain the probability of a verb selecting for a class, rather than the strength of this association. This is calculated by assigning probabilities to the networks based on some intuitions about how likely it is to select a class in relation to how likely it is to select its parent classes. See section B.3.7 for further details.

A class-to-class model Agirre and Martinez (2001) present a model for class-to-class preference acquisition, where preferences for classes of verbs - rather than preferences for a single verb - are learned. This strategy has two main justifications: one is that different senses of a verb can correspond to different preferences, the other being that some classes of verbs may have the same preferences.

In their procedure, the word-sense annotation in the SemCor corpus (Miller et al., 1993) and the structure of WN are exploited to estimate the probabilities of finding a

[^28]WN class in a certain argument position for a verb and to extend this to all the verb's senses. See section B.3.8 for more details.

Task-driven systems for SP acquisition WN has also been used in task-specific models for SP acquisition.

In the context of automatic clustering of German verbs, Schulte im Walde (2004) extracts coarse SPs from top-level nodes in GermaNet (Hamp and Feldweg, 1997; Kunze, 2000), the German version of WN. Schulte im Walde's method consists of splitting the frequency of a verb-object pair among the different senses of the noun and then upwards among its hypernyms. She proves that adding SP information obtained this way improves the verb clustering performance.

Calvo and Gelbukh (2004) introduce an unsupervised method that identifies SPs for use in PP-attachment resolution. For each noun extracted in a sequence containing a verb or a noun and a preposition, they make use of their semantic classification(s) in the Spanish WN and the corpus frequencies to decide if a noun is attached with the preposition to the preceding noun or to the near verb.

Pantel et al. (2007) use SPs to decide about admissible instances of inferential rules ${ }^{7}$ by using two semantic classifications: the WN noun hierarchy cut at depth 4 and the output of a clustering algorithm.

### 5.2.2 Knowledge-free approaches

Knowledge-free approaches in computational semantics aim at collecting information on a word's semantics by looking at its context, following the famous remark by Firth (1957, p. 11)

You shall know a word by the company it keeps!
This has led to distributional approaches to semantics in computational linguistics.
In the case of SP acquisition, these approaches share the first phase of data collection with the WN-based ones. However, in the generalization phase, distributional methods produce semantic classes directly from the corpus evidence on the words' co-occurrence contexts, rather than by exploiting an external classification such as WN.

### 5.2.2.1 An example on birds and insects

For example, suppose we are interested in the subjects of $f l y$, for which we collected the fillers and frequencies listed in table 5.1. ${ }^{8}$

We may want to split this set of nouns into coherent subgroups; for this, we could decide to place similar nouns into the same group. For example, we may expect crow and eagle to be more similar than, say, crow and bee: the contexts of these nouns may be extracted as the verb lemmas with which they occur. This intuition would be distributionally confirmed by the data if crow and eagle shared more contexts than crow and bee: in this case we would say that they are distributionally similar and distributional similarity would be taken as a model for semantic similarity.

[^29]| noun | frequency |
| ---: | :---: |
| bee | 3 |
| bug | 2 |
| bird | 4 |
| crow | 2 |
| eagle | 4 |
| insect | 5 |
| swallow | 2 |

Table 5.1: Example corpus frequencies for the subjects of $f l y$.

| noun | with fly | with buzz |
| ---: | :---: | :---: |
| bee | 3 | 5 |
| bug | 2 | 2 |
| bird | 4 | 1 |
| crow | 2 | 0 |
| eagle | 4 | 0 |
| insect | 5 | 10 |
| swallow | 2 | 1 |

Table 5.2: Constructed corpus frequencies of the nouns in table 5.1 as subjects of $f l y$ and buzz.

Table 5.2 shows the nouns displayed in table 5.1 with frequencies as subjects of $f l y$ and buzz in our toy example. We can now identify these seven nouns in geometrical terms by their co-occurrence with the two verbs: this means representing them as points in a space defined by two dimensions, one associated with $f l y$ and one with $b u z z$, as shown in figure 5.2. ${ }^{9}$ From an inspection of figure 5.2 we notice that the points crow, swallow and bug lie very close together, just like bird and eagle: this is because of their absolute frequencies, from which we may conclude that the distributions of these points are similar.

If we instead plot their relative frequencies (table 5.3), ${ }^{10}$ we get a different picture. In figure 5.3 insect and swallow, and crow-eagle are now close together. In fact, we may be interested in highlighting that insect appears twice as often with buzz as with fly (that is, one third and two thirds of its total frequency, respectively): this allows us to notice the similarity between eagle and crow, in spite of their different raw frequencies.

### 5.2.2.2 Clustering

The procedure just suggested can be viewed as the starting point of a clustering, a data-analysis technique illustrated more extensively in section B.5. The comparison between figure 5.2 and figure 5.3 showed that the choice between absolute and relative

[^30]

Figure 5.2: Two-dimensional plot from table 5.2, representing the subjects of $f l y$ and buzz (absolute frequencies).


Figure 5.3: Two-dimensional plot from table 5.3, representing the subjects of $f l y$ and buzz (relative frequencies).

| noun | with fly | with buzz |
| ---: | :---: | :---: |
| bee | 0.38 | 0.63 |
| bug | 0.5 | 0.5 |
| bird | 0.8 | 0.2 |
| crow | 1 | 0 |
| eagle | 1 | 0 |
| insect | 0.33 | 0.66 |
| swallow | 0.66 | 0.33 |

Table 5.3: Relative frequencies of the nouns in table 5.1 as subjects of $f l y$ and buzz.
frequencies - as other factors mentioned in section B. 5 - may considerably affect the results of the grouping.

A clustering algorithm, such as an agglomerative hierarchical clustering, is able to organize a given set of observations - the seven nouns with their frequencies in the example illustrated in section 5.2.2.1 - into groups called clusters, according to specific settings, such as a distance measure and a clustering method. It then produces a tree showing the output clusters. For example, the tree in figure 5.4 displays the output of a clustering on the constructed data from the previous example, where the chosen distance is the cosine distance: the relative frequencies of the objects (figure 5.2) - rather than their absolute distances (figure 5.3) - play a role in calculating this distance. ${ }^{11}$


Figure 5.4: Dendrogram from a clustering for the data in table 5.2 on page 88.

### 5.2.2.3 Survey of the systems

Clustering has been widely used in various tasks in computational semantics that require an automatic classification of words: for example for the creation of classes of adjectives (McKeown and Hatzivassiloglou, 1993), nouns (Pereira et al., 1993) and verbs (Schulte im Walde, 2003).

[^31]Here, we are interested in uses of clustering and similarity measures in SP acquisition, for which an overview is given in this section.

Rooth et al. (1999a) Rooth et al. (1999a) propose to cluster both argument fillers and verbs according to specific SC frames. The key idea is to obtain a joint probability distribution $p(v, n)$ through $p(c, v, n)$ where $c$ is a "hidden" class. They use the expectation-maximisation algorithm (Dempster and Rubin, 1977) to find the best classes. See section B.6.1 for more details.

Erk (2007) Erk (2007) proposes a distributional method for SPs for resolving Semantic Role Labeling. Her method makes use of two corpora: a primary corpus and a generalization corpus. First, she extractes all triples $(v, r, n)$ from the primary corpus. Then, SPs for a noun are calculated on the basis of its similarity with all fillers of $v$ for $r$; this similarity is computed on the generalization corpus through several metrics widely used in distributional semantics. See section B.6.2 for a more detailed description of her approach.

Bamman and Crane (2008) Bamman and Crane (2008) report experiments on extracting SPs from a 3.5 million word Latin corpus which was automatically tagged and parsed using the LDT as training set. The large size allows for better variance estimates and more representative frequencies: this overcomes the problems caused by the noisy input data (see section 4.3.3.2, page 53 ). SPs are then extracted through the log likelihood test, a technique also used in collocation extraction. See section B.6.3 for the formulae.

### 5.3 Our model

The previous systems for SP acquisition described in sections 5.2.1 and 5.2.2 exploit large corpora as training sets. This increases the probability of capturing a relative large set of argument fillers from the input. In addition, the size of the corpora allows for the use of statistical techniques that are able to highlight significant differences in frequency between likely and unlikely verbal arguments.

Our dataset consists of an SC lexicon extracted from two treebanks whose sizes are rather small by computational standards: approximately 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 is particularly problematic in the generalization step, since the set of seen events we are trying to generalize from may be too small to reflect the actual SP behaviour of Latin verbs. It is reasonable to deduce that employing unsupervised techniques to these data results in poor performance.

We tested if this state of affairs improves when the observations are grouped together to form larger sets from which the generalization is drawn. This idea was proposed by Alishahi (2008). The originality of Alishahi's approach is in fact an incremental clustering algorithm for verb occurrences called frames and identified by specific syntactic and semantic features, such as the number of verbal arguments, the syntactic pattern, and the semantic properties of each argument. Based on a probabilistic measure of similarity between the frames' features, the clustering produces larger linguistically motivated sets called constructions, which combine syntactic and semantic features. Constructions for a verb form the set of observations for the generalization step in Alishahi's system, which acquires the verb's SPs.

As will be illustrated more in detail in section 5.3.2, we modified Alishahi's system to adapt it to the input data at hand. We decided to test if the clustering proposed by Alishahi was effective in our case to solve the small size problem. In addition, we further extended the set of observations from which the generalization was drawn as follows. We included in this larger set not only the verb for which we are calculating the SPs and its lexical fillers, but also its synonyms and their fillers. The contribution of the verb's synonyms is weighted based on their similarity with the verb and the similarity between their fillers and the fillers of these synonyms.

Alishahi's work aims at cognitively modeling children's early acquisition of argument structure of verbs. Consequently, the corpora are used as collections of usage utterances because her system models the way children learn the verbs' argument structure by combining their exposure to language with a semantic taxonomy (WN). We believe that using a cognitively plausible model that combines contributions of syntax and semantics to SPs was a good choice in our case too. In fact, as said in section 1.1.1.2, our methodological aim was to show how computational methods can be built and used for modeling semantic phenomena in Latin. In addition, our model also meets the other goal of this thesis: creating a lexical resource for Latin to be employed in corpus studies.

In what follows, we will make use of an example to give a picture of our system and illustrate the main concepts that allow the generalization over unseen instances.

The problem Let us consider the verb cado 'fall down', which appears six times in our lexicon with an accusative object introduced by the preposition in 'towards' to express a motion argument: we will call this argument position '(in)Obj[acc]'. ${ }^{12}$ The corresponding indirect objects of cado are lexically instantiated in our lexicon as:

- culpa 'fault',
- esse 'being',
- fons 'spring, source',
- forma 'form',
- pars 'part',
- terra 'earth'.

Note that here and in the rest of this work we will not include pronouns among the lexical fillers of verbal arguments, but only nouns. In this, we follow the standard procedure in SP acquisition. This choice is usually explained by the assumption that it is this PoS that usually appears in argument positions for which SPs are searched for. However, as we will notice later (see table 5.13), a large number of argument slots in our corpora are filled by pronouns. A semantic classification of the pronouns' referents would need a manual annotation or an anaphora resolution step: due to the size of the data, neither was feasible in our case.

According to the definition of SPs given in chapter 1, we aim at answering the following question:

How can we generalize over the corpus occurrences in order to describe the preferences of the verb cado towards its arguments for '(in) Obj[acc]'?

[^32]| terra/11 | LWN <br> English | \{caelum, globus, humus, mundus, sphaera, tellus, terra\} <br> \{Earth, world, globe $\}$ |
| ---: | ---: | :--- |
| fons/3 | LWN | \{fons \} |
|  | English | \{spring, fountain, outflow, outpouring, natural_spring $\}$ |

Table 5.4: Synset number 11 for terra 'earth' and synset number 3 for fons 'source' in LWN, with their corresponding English synsets.

This question can be further refined into two different kinds of questions:
1 which nouns other than the lexical fillers listed above are likely to be found in the '(in) Obj[acc]' argument position for cado?

2 what are the most likely semantic properties associated with that argument position for cado?

WN similarity As we saw in section 5.2, question 1 is connected with word-based approaches and concerns grouping together similar nouns that are fillers of a given argument position. For example, we could ask what would be the chances of finding crimen 'verdict', 'crime', globus 'globe' or aqua 'water' following in as arguments of cado. As we will see, our system answers this question by considering the similarity in LWN between terra, fons, culpa and crimen, globus aqua.

Question 2 is identified by class-based approaches and concerns making the semantic features shared by the observed argument fillers explicit, either by using an external ontology - such as WN - that provides the semantic classes, or by inducing these classes in an unsupervised way through distributional methods.

In WN terms, if we move from the lexical to the conceptual level, we can see each lemma as associated with as many concepts as its different senses (synsets). Each of these synsets is embedded in higher-level synsets (hypernyms) forming a path along WN from the leaf (synset of the lemma) to the root node(s). Our system answers question 2 by considering the similarity between these paths. For example, terra belongs to 14 WN synsets and fons to 3 synsets; one synset for each lemma is illustrated in table 5.4. Figure 5.5 shows that the synset number 11 for terra shares a common path with the third synset for fons: the items contained in these hypernym synset paths


Figure 5.5: Portion of the LWN noun hierarchy containing the nodes 'terra/11' (the 11th synset for terra) and 'fons/3' (the third synset for fons).
as 'semantic properties', then we can state that terra/11 and fons/3 share are the semantic properties OBJECT-PHYSICAL OBJECT and ENTITY-SOMETHING.

Generalization If we had a larger set of fillers to generalize from, the approximation of the SPs for cado we just suggested would probably improve.

Let us consider a synonym of cado, namely descendo, ${ }^{13}$ which occurs with the following fillers for '(in) $\mathrm{Obj}[\mathrm{acc}]$ ' in our lexicon:

- homo 'man',
- solium 'tub', 'throne',
- terra 'earth'.

If we join the set of fillers for cado listed on page 92 with this set, we get a wider semantic range, which includes abstract nouns - culpa 'fault', esse 'being', forma 'form', pars 'part' - as well concrete nouns, like fons 'spring, source', and another instance of terra 'earth'. This gives additional reasons to infer a stronger preference of cado for terra and similar words, as well as more information on non-figurative usages of this verb, which we can record in the form of SPs towards concrete nouns. This hypothesis is confirmed by figures 5.6 and 5.7 , which show the portion of the WN hierarchy for the fillers of cado and descendo, respectively. ${ }^{14}$ Note that the filler terra is shared by the two verbs but only appears in figure 5.6 for reasons of space.

Suppose that we are interested in the top WN nodes as a coarse generalization over all the nodes associated to a verb's fillers. Figure 5.6 shows that the following top nodes generalize over the fillers for cado, except terra and forma: state, event, ABSTRACTION, ACT and PSYCHOLOGICAL-FEATURE. If we want to account for the broadest generalization for all senses of the fillers of cado - including terra and forma we have to add the top nodes Entity-SOMETHING, GROUP-GROUPING and POSSESSION to state, event, abstraction, act and psychological-Feature. This set covers all the fillers for descendo, too.

As we will explain in section 5.3.2.1, our system is able to group together some of the verbal corpus occurrences of cado and descendo, by measuring the semantic similarity between the two verbs, the closeness between their syntactic patterns containing '(in) $\mathrm{Obj}[\mathrm{acc}]$ ', and the semantic similarity between their fillers for this slot. ${ }^{15}$ In other words, rather than stating a general similarity between ascendo and cado, the system detects those single verb usages where these two predicates behave similarly with respect to syntactic and lexical-semantic criteria. Once these occurrences of descendo and cado have been grouped together, they all contribute to the set from which the system generalizes to create the SPs for descendo.

[^33]

Figure 5.7: Portion of the LWN noun hierarchy containing the fillers of descendo in our corpus: homo 'man', solium 'throne' and terra 'earth'.

A probabilistic model The main idea that drives Alishahi's and our model is that verbal SPs are viewed as probability distributions over all the possible semantic properties. They are calculated from corpus frequency information joined with the structuring of verb occurrences into constructions and with the architecture of WN. Finally, the constructions themselves are probabilistically defined as associations between syntactic and semantic features.

According to this reasoning, our system is able to calculate:

- the probability that a specific lemma contained in LWN is an argument for that verb;
- the probability that a specific semantic property is an argument for that verb.

These probabilities constitute answers to questions 1 and 2 on page 93 .

### 5.3.1 Overview

Alishahi (2008) proposes a computational model for SP induction in a cognitive framework. In this model each verb usage is defined as a frame, that is a collection of syntactic and semantic features, such as the number of verbal arguments, the syntactic pattern, and the semantic properties of each argument. A construction is defined as a collection of frames probabilistically sharing some syntactic and semantic feature values, for example the transitive construction. SPs are then calculated as probability distributions over the set of semantic properties, i. e. the set of WN nodes.

We partially modified the central definitions of frame and construction in Alishahi's model, as well as the formulae for the probability distributions. The present section addresses these details, in order to match the characteristics of our input data.

### 5.3.1.1 Frames and constructions

As chapter 4 illustrates, in our SC lexicon each verbal form occurring in the treebanks corresponds to a lexical entry recording syntactic, morphological and lexical information on the verb's arguments.
(1) grando magna sicut talentum descendit de hail:NOM.F.SG big:NOM.F.SG as talentum:NOM.M.SG come down:IND.PF.3SG from caelo in homines (Jerome, Apocalypse, 16.)
sky:ABL.N.SG towards man:ACC.M.PL
'Hailstones, as big as a talent, came down from the sky on men.'
Sentence (1) represents an active occurrence of the verb descendo 'go down' and is recorded in the lexicon as the following SC structure: ${ }^{16}$

$$
\begin{equation*}
\text { descendo }+ \text { A_(de)Obj[abl]\{caelum\},(in)Obj[acc]\{homo\},Sb[nom]\{grando\} } \tag{2}
\end{equation*}
$$

Each such verbal occurrence observed in the treebanks and represented in the form exemplified in (2) was defined as frame. Frames correspond to the utterances employed by Alishahi (2008). By associating frames to semantic features from an external ontology (LWN), we assume that they implicitly contain meaning elements, which we are going to make explicit in our model and illustrate in the next sections. Table 5.5 shows

[^34]

Table 5.5: The frame for sentence (2).
an example of the frame corresponding to sentence (2): ${ }^{17}$ each frame feature will be described more in detail in the next paragraphs.

Morpho-syntactic feature: feature $_{1}$ The argument positions for the active form of descendo in sentence (1) are an ablative object with preposition 'de' (A_(de)Obj[abl]), an accusative object with preposition 'in' (A_(in)Obj[acc]), and a nominative subject (A_Sb[nom]). The lexical fillers occurring in these positions are respectively: caelum 'sky', homo 'man', and grando 'hail'.

If we focus on the morpho-syntactic labels, we isolate the first feature of the frame: feature $_{1}$. The value of this feature in sentence (2) is:
A_(de)Obj[abl],(in)Obj[acc],Sb

In Alishahi's model the syntactic features are the number of verbal arguments in the frame, the participant roles of the arguments and the syntactic pattern. Since our input is not provided with the participant roles of the arguments and we do have their syntactic and morphological labels instead, we decided to include these labels in the definition of feature ${ }_{1}$.

Given the high optionality of verbal arguments in Latin and the fact that Latin is a Pro-Drop language,,$^{18}$ we excluded the number of arguments from the list of the frames' syntactic features.

Finally, the annotation guidelines of the treebanks (Bamman et al., 2008) prescribe that a noun in an ablative absolute construction is annotated as subject (cf. section 4.5.2.2); in these cases we find ablative nouns annotated as subjects, which, for our purposes, do not differ from nominative subjects. For this reason, we did not record the case marking of subject slots. Similarly, we did not account for the case marking

[^35]the argument labels 'OComp' (object complement) and 'Pnom' (predicate nominal). The case of objects ('Obj') is of course different, since the case marked by an object differentiates patterns of valency in Latin.

Semantic features of the arguments: feature $_{2}$ As in Alishahi's model, in addition to the morpho-syntactic features the lexical fillers of the verbal arguments are included into each frame (together with their WN mappings) as a semantic component. This is feature ${ }_{2}$.

For each argument position, the values of feature $_{2}$ are the set of WN synonyms and hypernyms of the filler(s) for that slot. For example, the semantic properties for the slot A_Sb in (2) are all the hypernyms of the synset for grando 'hail', as listed in figure 5.8.


Figure 5.8: Portion of the LWN noun hierarchy containing the semantic properties of grando 'hail'.

As described in section 3.3.1, the LWN database contains 9048 lemmas ${ }^{19}$ aligned with the English Princeton WN (PWN). The mapping links each Latin synset to an English synset and defines a "lexical gap" when this is not possible. Since LWN only records nouns, verbs, adjectives and adverbs, we labeled all occurrences of pronouns as argument lexical fillers in the corpora with the tag ' PR ': these fillers were not mapped onto LWN and were excluded from the next algorithmic procedures. In addition, we tagged all verbs and particles occurring as argument fillers with the labels ' V ' and ' Pt ' respectively. Finally, proper nouns can be all grouped into the same semantic category for SP purposes. For this reason, we manually extracted all the named entities from the list of fillers and tagged them as follows: 'NP' for names of person and people, like Achilles and Allobroges (anthroponyms and ethnonyms), 'NL' for names of place like Roma (toponyms), 'NO' for names of literary works (e.g. Ilias), and 'NT' for names of time entities, e. g. Pentecoste. ${ }^{20}$

In order to define a correct mapping between the lexical fillers extracted from the treebanks and the LWN lemmas, and between the lemmas in the LDT and those in the IT-TB, we performed a normalisation of the spelling mismatches.

[^36]| LWN | LDT | IT-TB | chosen |
| :---: | :---: | :---: | :---: |
| frons/1,7,8 | frons | frons ${ }^{\wedge}$ frondis | frons |
| frons/2,3,4,5,6 | frons2 | frons^frontis | frons2 |
| accido $/ 1,2,6,7,8,9,10$ | accido | accido^caedo | accido |
| accido/3, $4,5,11,12,13,14,15,16$ | accido2 | accido^cado | accido2 |
| impedio, impello | inpello | impedio | impedio, impello |
| adfligo | affligo | affigo | affligo |
| iota | adiungo | adjungo | iota, adiungo |

Table 5.6: Spelling differences between LWN synsets and lemmas in LDT and in IT-TB. The last column reports the standard we chose. For the LWN lemmas, we indicate the synset number after the '/' sign.

| descendo/1 | LWN | PWN |
| :--- | ---: | :--- | | \{cado, defluo, degredior, delabor, demeo $\}$ |
| :--- |
| \{descend, fall, go down, come down $\}$ |

Table 5.7: Synset for descendo 'come down' in our enriched version of LWN, with its corresponding English synset in PWN.

The main categories of these mismatches are listed in table 5.6 and concern:

- lemma homonymy: e.g. frons and accido; ${ }^{21}$
- spelling variations with or without assimilation of the consonant: e.g. impedio/impello and adfligo/affligo; ${ }^{22}$
- spelling of the semi-vowel $i$ as $j$ : e.g. adiungo vs. adjungo.

We opted for the spellings marked in the final column of table 5.6.
This phase gave us the opportunity to detect and correct some obvious mistakes in the lemmatization of the treebanks.

Semantic features of the verb: feature $_{3}$ Within the general purpose of acquiring the verbs' argument structure in children language, Alishahi's system generates the verbs' primitives, i.e. their 'event-specific meanings' (Alishahi, 2008, p. 16), such as <cause, move> for put. However, our system simply aims at acquiring SPs and we do not have a semantically annotated corpus; hence, we decided to use the information on the verb's semantics provided by LWN.

Consequently, the semantic feature of the verb belonging to a frame (feature $e_{3}$ ) is defined as the list of the verb's LWN synsets. In sentence (2) they coincide with the list of all items appearing in the synsets of descendo, displayed in table 5.7. ${ }^{23}$

[^37]Enriching Latin WordNet The coverage of LWN was low with respect to our data: 1027 fillers out of 2934 and 90 verbs out of 559 were not present in the lexical database.

Previous research on extending the WN ontology has faced this problem from different points of view and proposed solutions to automatically insert various kinds of information into WN: domain, contextual, WSD-oriented, etc. See, for example, Agirre et al. (2000), Harabagiu and Moldovan (2000), Navigli et al. (2004), Bentivogli and Pianta (2004).

In order to increase the number of LWN lemmas matching the fillers, we created a procedure for semi-automatically adding new Latin lemmas to the LWN hierarchy.

1 For each lemma $L$, we collected its Italian and/or English translations $T$ by using electronic versions of Latin-to-Italian and Latin-to-English dictionaries. For example, the lemma ianua 'door' does not appear in LWN; its Italian and English translations in the dictionaries Lewis and Short (1879) and Castiglioni and Mariotti (1993) are: porta, uscio, entrata, door, entrance.

2 Then, we extracted the synsets of these translations from the Italian version of MultiWordNet (MWN) and from the PWN. Table 5.8 shows the synsets for porta, entrata, door and entrance, with the domain label and the English or Italian glosses. ${ }^{24}$

3 In accordance with the semantics of the Latin word $L$, we manually selected the synsets of $T$ that are relevant to the senses of $L$, among all the possible senses. In the previous example they are 'porta/1', 'porta/2', 'porta/6', 'entrata/2', 'door/1', 'door $/ 3$ ', 'entrance $/ 1$ ' and 'entrance $/ 3$ ' were selected.

4 Thanks to the alignment between MWN, LWN and PWN, we were able to retrieve the existing Latin synsets corresponding to the selected Italian and English synsets from the previous step, when they were present; we finally assigned $L$ to these Latin synsets. For ianua, table 5.9 contains the LWN synsets corresponding to the selected senses of its translations. Therefore, we assigned the lemma ianua to the following synsets: \{accessus, aditus, foris, ingressus, initus, introitus, limen, porta, vestibulum \}, \{foris, porta\} and \{aditus, foris, ingressus, initus, introitus, limen, porta, vestibulum $\}$.

This procedure allowed us to add 401 new noun lemmas and 90 verb lemmas to 2056 already existing Latin synsets.

The manual selection of the relevant senses (step 3 above) caused the exclusion of 25 synsets only. In fact, most of the non-relevant synsets did not have a Latin counterpart, like 'porta/3'. This means that if the method was extended to a larger scale in an automatic way and without a manual intervention (step 3), the percentage of error would probably be small (error rate: $25 /(2056+25)=1,2 \%)$.

### 5.3.2 Selectional preference acquisition

In this section we will illustrate the two modules of our system: the Bayesian clustering of frames and the probabilistic model for SP acquisition.

[^38]| sense | domain |
| :--- | :--- |
| porta/1 | Buildings |
| porta/2 | Builidings |
| porta/3 | Sport |
| porta/4 | Skiing |
| porta/5 | Factotum |
| porta/6 | Factotum |
| porta/7 | Computer |
| porta/8 | Geology |
| entrata/1 | Factotum |
| entrata/2/2 | Transport |
| entrata/3 | Factotum |
| entrata//4 | Money |
| entrata/5 | Buildings |
| entrata//6 | Factotum |
| entrata//7 | Money |
| door/1 | Buildings |
| door/2 | Builingss |
| door/3 | Factotum |
| door/4 | Vehicles |
| door/5 | Buildings |
| door/6 | Buildings |
| entrance/1 | Transport |
| entrance/2 | Factotum |
| entrance $/ 3$ | Factotum | Table 5.8: Select

a swinging

| sense | domain | gloss | synset |
| :---: | :---: | :---: | :---: |
| porta/1 | Buildings | a swinging or sliding barrier that will close the entrance to a room or building | \{porta\} |
| porta/2 | Buildings | the space in a wall through which you enter or leave a room or building; the space that a door can close | \{porta\} |
| porta/3 | Sport | a place toward which players of a game try to advance a ball or puck in order to score points | \{rete, porta, bersaglio \} |
| porta/4 | Skiing | nello sci, paletto o coppia di paletti indicanti il passaggio obbligato per i concorrenti nelle gare di slalom | \{porta\} |
| porta/5 | Factotum | elemento di chiusura o apertura di un vano o di veicoli | \{porta\} |
| porta/6 | Factotum | anything providing a means of access (or escape) | \{porta\} |
| porta/7 | Computer science | hardware and associated circuitry that links one device with another (especially a computer and a hard disk drive or other peripherals) | \{porta, interfaccia\} |
| porta/8 | Geology | the location in a range of mountains of a geological formation that is lower than the surrounding peaks | \{passo, valico, colle, porta\} |
| entrata/1 | Factotum | the act of entering | \{entrata, ingresso\} |
| entrata/2 | Transport | something that allows entry or exit | \{accesso, adito, entrata, ingresso\} |
| entrata/3 | Factotum | the act of beginning something new | \{avviamento, debutto, esordio, instaurazione, istaurazione, lancio, entrata\} |
| entrata/4 | Money | the entire amount of income before any deductions are made | \{entrate, incasso, introito, entrata \} |
| entrata/5 | Buildings | a large entrance or reception room or area | \{androne, anticamera, atrio, foyer, hall, ingresso, ridotto, vestibolo, entrata\} |
| entrata/6 | Factotum | the act of admitting someone to enter | \{accesso, ammissione, entrata\} |
| entrata/7 | Money | the financial gain (earned or unearned) accruing over a given period of time | \{cespite, entrata, introito, reddito, utile\} |
| door/1 | Buildings | a swinging or sliding barrier that will close the entrance to a room or building | \{door\} |
| door $/ 2$ | Buildings | the space in a wall through which you enter or leave a room or building; the space that a door can close | \{doorway, door, room access, threshold\} |
| door $/ 3$ | Factotum | anything providing a means of access (or escape) | \{door\} |
| door $/ 4$ | Vehicles | a swinging or sliding barrier that will close off access into a car | \{door\} |
| door $/ 5$ | Buildings | a house that is entered via a door | \{door\} |
| door/6 | Buildings | a room that is entered via a door | \{door\} |
| entrance/1 | Transport | something that allows entry or exit | \{entrance, entranceway, entryway, entry, entree\} |
| entrance/2 | Factotum | a movement into or inward | \{entrance, entering\} |
| entrance/3 | Factotum | the act of entering | \{entrance, entering, entry, ingress, incoming\} |

Table 5.8: Selected synsets for porta, entrata, door and entrance in MWN, with their domain and glosses.
$\square$

| Italian synset | Latin synset |
| ---: | :--- |
| porta $/ 1$ | \{foris, porta\} |
| porta $/ 2$ | none |
| porta $/ 6$ | none |
| entrata $/ 2$ | \{accessus, aditus, foris, ingressus, initus, introitus, limen, porta, |
|  | vestibulum |
| door $/ 1$ | \{foris, porta $\}$ |
| door $/ 3$ | none |
| entrance $/ 1$ | \{accessus, aditus, foris, ingressus, initus, introitus, limen, porta, |
|  | vestibulum $\}$ |
| entrance $/ 3$ | \{aditus, foris, ingressus, initus, introitus, limen, porta, vestibulum $\}$ |

Table 5.9: Synsets for porta, entrata, door and entrance in MWN.

### 5.3.2.1 Bayesian clustering of frames

In section 5.3.1.1 we defined a frame as a pattern identified by the values of three following features:

- feature $_{1}$ (morpho-syntactic): verb lemma + voice + syntactic and morphological tags of its arguments: e.g. 'descendo $+\mathrm{A}_{-}$(de) Obj[abl]';
- feature $_{2}$ (semantic): WN synonyms and hypernyms (semantic properties) of the lexical fillers of the verb's arguments: e. g. ATMOSPHERIC PHENOMENON, ENTITY;
- feature $e_{3}$ (semantic): WN synset of the verb: e.g. cado, descendo.

In order to generalize over the verbal occurrences in corpora, we aim at clustering such frames into constructions. A construction is here intended as a collection of frames that are similar from the point of view of their syntactic and semantic features. The required degree of this similarity will be explained in this section by specific similarity scores and probability formulae: these scores measure how many values of each feature are shared by a set of frames. For example, the frame (3) is clustered together with the frame (4)

```
descendo + A_(de)Obj[abl]{caelum},(in)Obj[acc]{homo},Sb{grando}
descendo + A_(de)Obj[abl]{caelum},(in)Obj[acc]{terra}
```

on the basis of the similarity between:

- the syntactic patterns of the two frames;
- the number of shared semantic properties of the two pairs of fillers for the slot '(de)Obj[abl]' (caelum) and the slot '(in)Obj[acc]' (homo and terra);
- the similarity/identity between the verbs of the frames: descendo.

Table 5.10 is an overview of the algorithm, which consists of the unsupervised incremental Bayesian process from Alishahi (2008).

1 The input is composed by the set of all $N$ frames observed in the corpora, each of which forms a construction containing just the frame itself.

2 Then, the first frame $F_{1}$ is taken into consideration and the similarity scores between $F_{1}$ and all other frames (i. e. constructions) are calculated. These scores contribute to the probability $P\left(k \mid F_{1}\right)$, which is the conditional probability of $k$ given $F_{1}$. See section B.1.1 for a definition of 'conditional probability'.

|  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| step 0 | $\left\{F_{1}\right\}$ | $\left\{F_{2}\right\}$ | $\ldots$ | $\left\{F_{x}\right\}$ | $\ldots$ | $\left\{F_{N}\right\}$ |
| step 1 | $\emptyset$ | $\left\{F_{2}\right\}$ | $\ldots$ | $\left\{F_{1}, F_{x}\right\}$ | $\cdots$ | $\left\{F_{N}\right\}$ |
| $\cdots$ |  |  |  | $\cdots$ |  |  |

Table 5.10: Schema of a clustering algorithm.

3 If we find a construction $K=\left\{F_{x}\right\}$ for which the probability $P\left(k \mid F_{1}\right)$ is maximised and greater than the baseline probability $P\left(k_{0} \mid F_{1}\right),{ }^{25}$ then $F_{1}$ is assigned to $K$, which then consists of two frames: $F_{x}$ and $F_{1} . K$ contains the frame that is the most similar to $F_{1}$ according to feature ${ }_{1}$, feature $_{2}$ and feature $_{3}$, as measured by the similarity scores. If such a construction cannot be found, this means that no frame is sufficiently similar to $F_{1}$, which is then assigned the singleton construction $\left\{F_{1}\right\}$.
4 In a similar way, the second frame $F_{2}$ is compared with all other constructions and assigned the most similar construction according to the similarity scores.

5 The algorithm keeps on considering a frame at a time until all frames are assigned to a construction.

This way, the membership of constructions progressively increases as new frames are added: at the end of the process, the set of constructions is determined as the output of the algorithm.

As can be noticed from the sequence of steps we just listed, the final output of the algorithm depends in theory on the order by which frames are assigned constructions. In fact, a different ordering that starts with, say, $F_{3}$ might give rise to a potentially different set of constructions. Nevertheless, the final calculation of SPs can be considered to be not affected. We tested this by randomly extracting 10 frames from our dataset; then the clustering algorithm was run 10 times by starting from each of the 10 frames. The outputs from these 10 iterations were used in the subsequent calculation of SPs, described in section 5.3.2.3. Finally, the probabilities $P(s \mid v)$ for each iteration were extracted, for each pair $(v, s)$ where $s$ is a semantic property extracted from a subset of 10 randomly chosen semantic properties and $v$ is a verb occurring in the test dataset. The standard deviation and the mean of these probabilities over the 10 iterations showed that the probabilities are not significantly different from each other.

We will now illustrate the main concepts behind the clustering algorithm, referring to section B. 8 for the mathematical details. We wrote a Perl script that implemented the algorithm.

A frame $F$ is assigned a construction $K$ if $K$ maximises the probability $P(k \mid F)$ over all constructions $k$. After Bayes' theorem, ${ }^{26}$ this is equivalent to maximising the product

$$
P(k) P(F \mid k) .
$$

The prior probability $P(k)$ measures how much $k$ is "attested" in the corpora and is calculated as the number of frames contained in $k$ divided by the total number of frames.

[^39]On the other hand, $P(F \mid k)$ is calculated by examining the degree of syntactic and semantic similarity between the frame $F$ and the frames already included in the construction $k$. In other words, $P(F \mid k)$ calculates the overlap between the values of the features (feature ${ }_{1}$, feature $e_{2}$ and feature $e_{3}$ ) in $F$ and the values of the same features in the other frames belonging to $k$. If we assume that the frame features are independent, $P(F \mid k)$ is the product of the conditional probability that each feature displays the same value it has in $F$ and in $k$. We will now describe these probabilities.

Feature $_{1}$ For the syntactic properties, we estimated the conditional probability of finding the same values for feature $_{1}$ in $F$ and in $k$ by calculating the sum of the syntactic scores between $F$ and each frame $h$ contained in $k$.

The syntactic score is the number of syntactic slots shared by $h$ and $F$ divided by the number of slots in $F$. This accounts for the degree to which the syntactic patterns of $h$ differ from those of $F$. Given the high frequency of omitted arguments in Latin sentences, the chances of an exact match between the two SCSs are low. For this reason, we introduced this score instead of simply counting the number of frames in $k$ that share the exact same syntactic patterns with $F$, as Alishahi (2008) does. For example, let us consider the frame given by (3) as $F$ and the frame (4) as $h$ (page 102). The syntactic score between $F$ and $h$ is given by the number of slots shared by $F$ and $h$ (which is 2: 'A_(de)Obj[abl]' and 'A_(in)Obj[acc]'), divided by the total number of slots in $F$ (which is 3 ), hence it is $\frac{2}{3}$.

Feature $_{2}$ For the semantic properties of the argument fillers, we calculated the sum of semantic scores.

The semantic score accounts for the degree of overlap between the semantic properties of the argument fillers in $F$ and the semantic properties of the fillers in each frame contained in $k$; it then divides this overlap by the number of semantic properties of the fillers in $F$. In the frames represented in sentences (3) and (4) (which we will call frame $F$ and frame $h$, respectively), terra 'earth' and homo 'man' are the lexical fillers of the '(in) $\mathrm{Obj}[\mathrm{acc}]$ ' slot for $h$ and $F$, respectively. The semantic score relative to the slot '(in) $\mathrm{Obj}[\mathrm{acc}]$ ' is $\frac{5}{175}=0.03$, because the intersection between the semantic properties of the two words contains 5 items (caelum, classis, mundus, natio and res) out of 58 (which are all the semantic properties for terra). ${ }^{27}$

Feature $_{3}$ For the semantic properties of the verb, we calculated the synset scores between the verb in $F$ and the verbs of the frames in $k$.

The synset score calculates the degree of overlap between the synsets for the verb in $F$ and the synsets for the verbs in each frame of $k$; it divides it by the number of synsets for the verb in $F$. The final probability is anyway never 0 because we used a smoothing technique. For example, let us compare the frame $F$ from sentence (3) with the following frame $h$ :

$$
\begin{equation*}
\text { cado+ A_Sb\{angelus }\} \tag{5}
\end{equation*}
$$

The synsets for cado 'fall' and descendo 'descend' are represented in table 5.11. The overlap between the synsets for descendo and those for cado contains one synset: \{cado, defluo, degredior, delabor, demeo $\}$. The number of synsets for $F$ is 1 , so the synset score is $\frac{1}{1}=1$. We introduced the syntactic and synset scores in order to account

[^40]| cado/1 | \{cado, defluo, degredior, delabor, demeo $\}$ <br> \{descend, fall, go down, come down |
| ---: | :--- |
| cado/2 | \{abscedo, degredior, migro, proficiscor\} <br> \{depart, part, start, start out, set forth, set off, set out, take off $\}$ |
| cado/3 | \{averto, declino, deflecto, degredior, derivo, detorqueo, deverto <br> \{deviate, divert $\}$ |
| cado/4 | \{degredior\} <br> \{stoop, descend $\}$ |
| cado/5 | \{degredior\} <br> \{condescend, deign, descend $\}$ |
| cado/6 | \{degredior, delabor\} <br> \{derive, come, descend $\}$ |
| descendo $/ 1$ | \{cado, defluo, degredior, delabor, demeo $\}$ <br> \{descend, fall, go down, come down $\}$ |

Table 5.11: Synsets of cado 'fall' and descendo 'descend'. For each sense of these two verbs, the first row contains the LWN synset, while the second row contains their corresponding English synsets in MWN.
for a frequent phenomenon in our data: the partial matches between the values of the features in $F$ and in $k$.

### 5.3.2.2 Output of the clustering

In this section we will provide some quantitative data on the output of the algorithm and a qualitative interpretation of the clustering results as well as an error analysis.

We ran the algorithm on the dataset containing the two joint sublexicons from LDT and IT-TB described in chapter 4. This corresponds to 11478 distinct frames with a total frequency of 15509. The number of distinct syntactic patterns (such as 'A_(ex)Obj[abl], Obj[acc]') is 697 and the distinct pairs (verb, syntactic pattern) are 5309. The output of the algorithm consists of 7105 constructions.

For example, the 5 frames where the verb introduco 'bring in, introduce, insert' occurs (frequency $=9$ ) are split into 5 different constructions, listed in table 5.12. Note the semantic similarity between the three verbs involved in construction 1, for example: induco 'bring forward, introduce', introduco and addo 'add to, bring to'. Similarly, introduco with induco are combined by the fillers of the PP introduced by in 'in' finis 'end', 'limit', 'borders' and effectus 'execution', 'accomplishment', 'defect', which share the semantic properties ATTRIBUTE, COGNITIO 'cognition', CONSCIENTIA 'conscience', EVENTUM 'event', among others. Note also the similarity between the fillers in construction 4: forma 'form' and perfectio 'perfection'.

Among the constructions acquired, 3011 contain one frame, 3836 contain 2 frames, 242 contain 3 frames, 13 contain 4 frames, 2 contain 5 frames and one contrains 6 frames. The small number of frame per construction and, consequently, the high number of constructions can be explained by the coverage of LWN 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 is 0 , which causes the probability $P(k \mid F)$ to be mostly lower than the threshold $P\left(k_{0} \mid F\right)$ (but not 0 because of the smoothing); this results in assigning the frame to a singleton construction consisting of the frame itself.

Furthermore, a large number of frames - 6588 out of $19581,33 \%$ - contain verbs,

| 1 | induco + P_Sb[acc]\{forma\} |
| :---: | :---: |
|  | introduco + P_Sb\{PR\} |
|  | introduco + P_Sb\{forma\} |
|  | addo +P _Sb praesidium $^{\text {a }}$ |
| 2 | induco + A_Obj[acc]\{forma\} |
|  | immitto + A_Obj[acc]\{PR\},A_Obj[dat]\{antrum\} |
|  | introduco + A_Obj[acc]\{NP\} |
| 3 | introduco + A_(in)Obj[acc]\{finis\}, A_Obj[acc]\{copia\}, A_Sb\{NP\} |
|  | induco + A_(in)Obj[acc]\{effectus\},A_Obj[acc]\{forma\} |
| 4 | introduco + A_Obj[acc]\{forma\} |
|  | induco + A_Obj[acc]\{perfectio\},A_Sb[nom]\{PR\} |
| 5 | induco + A_Obj[acc]\{forma\} |
|  | immitto + A_Obj[acc]\{PR\},A_Obj[dat]\{antrum $\}$ |
|  | introduco + A_Obj[acc]\{NP\} |

Table 5.12: Constructions for the frames containing the verb introduco 'bring in'.

| category | frequency |
| ---: | :--- |
| nouns of literary works 'NO' | 2 |
| anthroponyms and ethnonyms 'NP' | 989 |
| toponyms 'NL' | 66 |
| pronouns ' PR ' | 3884 |
| participles 'Pt' | 464 |
| verbs'V' | 1646 |

Table 5.13: Different categories of fillers in our lexicon and their frequency.
participles, pronouns and proper nouns as their fillers; consequently, if they are clustered with other frames, they are probably clustered with frames containing fillers with the same, since this is the only case where the semantic score is not 0 . Table 5.13 summarises all these cases.

In addition, in our system more frequent senses do not contribute more than less frequent ones; instead, senses with longer paths to the root nodes contribute more than the others, because the number of their hypernyms is higher. This biases the actual frequency distributions of senses in corpora. As an example, take the two nouns felicitas and laetitia. According to Lewis and Short (1879), the first meaning of felicitas is 'fruitfulness, fertility', while the second is 'happiness, felicity', considered as the predominant signification; the third one is the deity Felicitas. For laetitia, Lewis and Short (1879) lists 'unrestrained joyfulness, gladness, pleasure, delight' and then the transferred sense of 'pleasing appearance, beauty, ragrace'. Although the semantics of these two nouns may appear quite similar, their semantic scores are lower than 0.5 , because the intersection between the semantic properties of felicitas and those of laetitia contains only 12 items, while the union contains 51 .

Finally, as we pointed out in section 5.1, the small size of our sample underestimates the variance. If we assume that our dataset is a sample from a larger population of words whose frequency follows a Zipf-like skewed distribution - as is typically the case with linguistic data - then a larger sample would capture more variation in the population.

A larger sample would contain types that are unseen in our sample, but would still have the most frequent type as the dominant one; nevertheless, a higher frequency would be assigned to hapaxes observed in our data. ${ }^{28}$ See, for example, Gale and Sampson (1995) for a discussion on these issues.

All these difficulties make the constructions defined in our dataset less general than those found by Alishahi (2008). This may also be caused by the fact that she uses more homogeneous corpora: the children language corpus Childes and the general language corpus BNC. However, the constructions we obtained represent a secondary result of our investigation, which mainly focusses on acquiring SPs.

### 5.3.2.3 A probabilistic model for SPs

We define SPs as probability distributions over the set of semantic properties, which are elements of WN nodes (cf. section 5.3.2.1): for example Entity or object. We wrote a Perl script that automatically calculates these SPs.

In order to generalize from the lexical fillers collected from our lexicon to semantic properties, we refer to the set of constructions defined by the clustering algorithm described in section 5.3.2.1, to test the actual impact of clustering in our case. In particular, the probability of a semantic property occurring in a certain argument position for a verb $v$ is calculated based on all constructions whose verbs appear in the same synset(s) as $v$. Some of these constructions may contain the frames for $v$ as well as frames whose verbs are different from $v$; this happens when these frames are clustered by the similarity scores, like introduco and induco 'introduce, insert, bring in' in table 5.12. In addition, the summed probabilities are weighted by the synset scores between $v$ and each verb in the construction, thus accounting for the semantic similarity between the verbs themselves. The contribution from the verbs sharing a synset with $v$ is an innovation from the system by Alishahi (2008) and was tuned to our dataset, where the low frequencies in many cases would not allow the generalization step.

If $s$ is a semantic property and $a$ an argument position for a verb $v$, we calculate the probability $P_{a}(s \mid v)$ as the sum of $P_{a}(s, k \mid v)$ over all constructions $k$ containing $v$ or a WN-synonym of $v$. By Bayes' rule, this leads us to calculate $P_{a}(s \mid k, v)$, which is obtained by following the principles illustrated in section 5.3.2.1 and described more in detail in section B.8.2.

In particular, we want to account for:
1 the degree by which the verbs in the constructions $k$ are similar to $v$;
2 the degree by which the fillers of the frames in $k$ are similar to $s$.
This is achieved by taking the counts of frames in $k$ and weighting them by a similiarity score, defined as the product of:

1 the synset score between $v$ and the verbs in each frame belonging to $k$;
2 the overlap between $s$ and the semantic properties of the fillers of the frames in $k$.

See section 5.3.2.1 for the definitions of synset score and semantic score.
If we sum these similarity scores over all frames in $k$, we divide this sum by the total frequency of the frames in $k$ that contain $v$ or its synonyms and finally normalise,

[^41]| 1 | $\text { includo+ A_(in)Obj[acc]\{equus\},A_Obj[acc]\{NP\},A_Sb\{NP\} }$ |
| :---: | :---: |
|  | includo+A_Obj[acc]\{NUM (numeral)\},A_Sb\{praeparatio\} |
| 2 | navigo+ A_(in)Obj[acc]\{locus\},A_Sb\{NP\} |
|  | infero + A_(in)Obj[acc]\{materia\} |
| 3 | addo+A_(in)Obj[acc]\{sententia\}, A_(ut)Obj[sbjv]\{V\} |
| 4 | induco+ A_(in)Obj[acc]\{materia\}, A_Obj[acc]\{forma\} |
|  | conicio + A_(in)Obj[acc]\{locus\},A_Obj[acc]\{mulier\},A_Obj[sbjv]\{V\} |
| 5 | intelligo +A_Obj[acc]\{dimensio\} |
|  | conicio +A_(in)Obj[acc]\{os\},A_Obj[acc]\{aqua\} |
| 67 | conicio + A_(in)Obj[acc]\{folium\}, A_Obj[acc]\{dominus\} |
|  |  |
|  | conicio + A_(in)Obj[acc]\{balneum\}, A_Obj[acc]\{PR\} |
| 8 | immitto +A_(in)Obj[acc]\{facies, , A_Obj[acc] \{calix\}, A_Sb\{NP\} |
|  |  |
| 9 | induco + A_(in) Obj[acc]\{materia\}, A_Obj[acc]\{forma\},A_Sb\{agens |
|  | induco + A_(in)Obj[acc]\{materia\}, A_Obj[acc]\{PR\} |
| 10 | induco + A_(in)Obj[acc]\{animus, , A_Obj[inf]\{V\}, A_Obj[inf]\{V\}, A_Sb\{NP\} |
|  | induco + A_Obj[inf $]\{\mathrm{V}\}$ |
| 11 | introduco +A_(in)Obj[acc]\{finis\},A_Obj[acc]\{copia\},A_Sb\{NP\} induco $+\mathrm{A}_{-}$(in) $\mathrm{Obj}[\mathrm{acc}]\{$ effectus $\}, \mathrm{A}_{2} \mathrm{Obj}[\mathrm{acc}]\{$ forma $\}$ |

Table 5.14: Constructions for the frames containing the synonyms of the verb introduco 'insert, bring in' in LWN.
we get a measure of how likely it is to find the semantic property $s$ in the argument position $a$ for $v$ in the construction $k: P_{a}(s \mid k, v)$. For the details, see Appendix B.8.2.

As an example, let us consider the SPs for introduco 'introduce, bring in' and let us look at its active accusative arguments introduced by in 'in'. The synonyms of introduco in LWN are: \{addo, conicio, immitto, includo, indo, induco, insero, inserto, intericio, interpono, intexo, introduco $\}$ (first synset), \{addo, conicio, immitto, indo, induco, insero, inserto, intericio, interpono, intexo, introduco, penetro $\}$ (second synset), and \{adhibeo, importo, indo, induco, infero, inserto, intericio, interpono, introduco, inveho \} (third synset). For adhibeo, importo, includo, infero, inveho and penetro the synset score with introduco is $\frac{1}{3}$, since they appear in one of the three synsets of introduco; for addo, conicio, immitto, insero, intexo the synset score is $\frac{2}{3}$ and finally for indo, induco, inserto, intericio, interpono it is $\frac{3}{3}=1$. The frames contained in the constructions for introduco are listed in table 5.12. In addition, the constructions for its synonyms are listed in table 5.14. ${ }^{29}$

We now want to calculate the probability of the semantic property BEING as an argument of introduco for 'A_(in) $\mathrm{Obj}[\mathrm{acc}]$ '. For each frame in these constructions we calculate the SP similarity score. For example, let us consider the frame $h$ as includo+ A_(in) $\operatorname{Obj}[\operatorname{acc}]\{$ equus $\}, \mathrm{A}_{-} \operatorname{Obj}[\operatorname{acc}]\{\mathrm{NP}\}, \mathrm{A}_{-} \mathrm{Sb}\{\mathrm{NP}\} .{ }^{30}$ As we saw, the synset score between includo and introduco is $\frac{1}{3}$. The number of fillers of 'A_(in) $\mathrm{Obj}[\mathrm{acc}]$ ' in $h$ is 1 : equus 'horse'. The semantic properties of horse are:
animal, animalis, apparatus, artefact, artifact, being, bellua, bestia, caballus, chess_piece, chess_piece, chessman, chordate, craniate, entity, eques, equid, equine, equus, eutherian, eutherian_mammal, game_equipment, hoofed_mammal, instrumentality, instrumentation, life_form, living_thing, mammal, man, object, odd,

[^42]organism, pecus, perissodactyl, perissodactyl_mammal, physical_object, piece, placental, placental_mammal, something, sonipes, toed_ungulate, ungulate, vertebrate.

Therefore, $h$ has the property BEING; thus the similarity score between $h$, introduco and BEING is $0.3 \cdot \frac{1}{1}=0.3$.

### 5.4 Evaluation

In this section we report on the evaluation procedures we conducted in order to assess the goodness of our model for SP acquisition. First, we discuss some properties of the SP probability distribution calculated by our model (section 5.4.1); further, we present the results of the evaluation we carried out on a sample set of verbs (section 5.4.2). Then, we give some details of a comparison between the knowledge-based version of our system for SP acquisition and a knowledge-free distributional version of the same system (section 5.4.3).

### 5.4.1 The probability distribution

The probabilistic model originally calculates $P_{a}(s \mid v)$ for a semantic property $s$; however, for the evaluation we concentrated on the probability $P_{a}(n \mid v)$, i. e. a modified version of the model. This is due to the fact that we work on corpus data, and we aim at testing the ability of our model to predict the probability of a specific noun $n$ occurring in a specific argument position for a verb. In order to obtain $P_{a}(n \mid v)$, we calculated the SP similarity score for the slot $a$ between frame $h$ and $(v, n)$ by changing the formula (B.13) on page 217, as explained in formula (B.14) on page 217: the effect is to calculate how similar $n$ is to the fillers of $v$.

Entropy As a first investigation, we compared our model with a baseline model where no previous information about a noun $n$ is exploited in order to guess its SP probability. This model assigns the uniform distribution over all nouns, i. e. $\frac{1}{4724}$ for each of the 4724 LWN leaf nodes. The entropy of this baseline model, as defined in section B.2, is 12.2. By the principle of Maximum Entropy (see again section B.2), this baseline model has the maximum entropy.

On the other hand, our model's entropy is always lower than the baseline: tested on a random set of 40 verbs, it ranges from 6,86 (intellego +A _Sb) to 11,52 (sum+A_Sb). This confirms that the system uses some information for estimating the probabilities. The information used is the WN structure and the position of verbs and fillers in LWN, as well as the co-occurrence frequencies and the syntactic-semantic content of the constructions defined by the Bayesian clustering.

Now, we will see how well the model uses this information by testing two of its properties: discrimination and calibration (Harrell et al., 1996, p. 366).

The model's calibration and discrimination Calibration is related to the model's ability to distinguish between high and low probabilities. Following the definition of SPs, we expect a small set of nouns (preferred nouns) to get a high SP probability, and a large set of nouns (the rest) to get a low probability.

As a way of example, figure 5.9 shows the SP distribution for the objects of dico 'say', as given by our model. The $y$ axis displays the SP probability values, while the $x$ axis displays the number of semantic properties assigned to a certain SP probability. In


Figure 5.9: Decreasing SP probabilities of the LWN leaf nodes for the objects of dico 'say'.
general the plots have shapes similar to figure 5.9: this shows that the SP distribution given by our model is very skewed for all verbs, with very few high-frequency nouns and a large number of low-frequency nouns. ${ }^{31}$ This result constrasts with the unrealistically flat graph of the baseline model and shows that our model is adequately calibrated.

Now, we test the system's discrimination potential, i. e. its ability to correctly estimate the SP probability of each single LWN leaf node. Table 5.15 displays the 15 nouns with the highest probabilities as direct objects of dico 'say'. As expected, from table 5.15 and figure 5.9 we see that the model assigns a high probability to most seen fillers for dico in the corpus: anima 'soul', corpus 'body', locus 'place', pars 'part', etc. In order to establish the model's discrimination, i.e. how accurate the obtained SP

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

Table 5.15: 15 nouns with the highest probabilities as accusative objects of dico 'say'.

[^43]probabilities are, we opted for a corpus investigation, as explained in section 5.4.2.

### 5.4.2 The evaluation set

The discrimination of many systems for SP acquisition has been evaluated on a specific task, such as PP-attachment Disambiguation, WSD or Semantic Role Labeling: cf. McCarthy (1997); Calvo and Gelbukh (2004); Erk (2007). Alishahi (2008) follows the evaluation approach by Resnik (1996), which involves the use of human plausibility judgements on verb-noun pairs.

Since we deal with a language with no native speakers (see discussion in sections 1.1.1.1 and 1.3.0.1), we approximated plausibility judgements with corpus occurrences. By simulating a random extraction of argument fillers for a list of verbs from a wide variety of Latin texts, we considered these instances as implicit plausibility judgements; in other words, we assumed that each such instance was considered acceptable by the speaker (writer).

We randomly selected a list of 40 verbs from our corpora (listed in table B. 12 on page 222), equally distributed in two classes of frequency:

1 high frequency ( $>51$ ): e.g. dico 'say',
2 medium frequency (between 11 and 50): e.g. ascendo 'ascend'.
For each of these verbs we selected one argument slot that was particularly interesting to look at. Then we manually collected a lexical filler for each such pair from the following of freely available collections of Latin texts: Perseus Digital Library, ${ }^{32}$ Corpus Thomisticum ${ }^{33}$, and The Latin Library. ${ }^{34}$ We excluded those fillers that already occurred in the training set (our SC lexicon). The list of fillers for each verb-slot pair constituted our gold standard. For each of the 40 verb-slot $(v, a)$ pairs, we focussed on the probability $P_{a}(f \mid v)$ assigned by the model to the random filler $f$ extracted.

Central tendency In order to determine how good this prediction was, we compared $P_{a}(f \mid v)$ with three measures of central tendency: mean, median and the value of the third quantile. The mean is the average value among all SP probabilities, the median was calculated by sorting the SP probabilities and picking the middle one; finally, the third quartile is the threshold under which three quarters of the nouns have their probabilities. See section B.9.0.1 for more detailed definitions and figure 5.10 for a visualisation of these measures for the direct objects of dico.

If the probability assigned by the model to the random filler was higher than the measures of central tendency, the outcome was considered a success. The successes were 28 for the mean, 36 for the median and 24 for the third quartile. This means that the probability of the random filler was higher than the mean in 28 cases out of 40 , and it was higher than the median in 36 cases. Finally, in 24 cases the random filler was placed in the high-probability quarter of the plot, which is where the preferred arguments gather. A binomial test found that the success rate was statistically significant at the $5 \%$ level for the first two measures and not statistically significant at the $5 \%$ level for the third measure. See section B.9.0.2 for the details of the test.

[^44]

Figure 5.10: Mean, median and value of the third quantile for the objects of dico 'say'.

Random filler vs. remaining 39 random fillers Next, we compared the probabilities assigned by the system to each verb's random filler with mean of the probabilities assigned to the 39 random fillers extracted for the other verbs. Our assumption was that in most cases the former probability should be higher than the latter (leading to a success), even though in some cases the random filler extracted for a verb-slot pair could be compatible with different verb-slot pairs. In a significantly higher proportion of cases ( 24 out of 40 ) we obtained a success, as confirmed by the test results given in section B.9.0.2 on page 223 .

The impact of clustering In order to evaluate what impact the clustering step actually had on the overall performance of our system, we ran a "clustering-free" version of the algorithm, where the constructions were defined as containing a single frame each. We then compared the results with the standard version of the algorithm. In 28 out of 40 cases the probability assigned to the random filler was higher than the mean of the probabilities; in 36 cases the probability was higher than the mean and in 24 cases the probability was higher than the third quartile. In addition, in 24 cases the probability assigned to the random filler for a verb-slot pair was higher than the mean of the probabilites assigned to the 39 random fillers for the remaining verb-slot pairs. These results coincide with the measures obtained for the standard version of the algorithm. The evaluation performed on fillers extracted from parsed corpora gave the same results on the clustering-free version and on the original version.

Overall, we can say that the clustering step did not affect the performance of the system, at least in relation to the measures we chose; this might be due to the high number of clusters discussed in section 5.3.2.2. Further research is thus needed to test whether the clustering step has an impact when is algorithm is trained on a larger dataset and with different smoothing factors.

### 5.4.2 . Fillers from parsed corpora

The approach to evaluation we just described does not take into account the corpus frequency of the random fillers. In other words, their plausibility as argument fillers for specific verbs is based on their single occurrences with those verbs, as observed in the random extraction process.

As an alternative approach to evaluation, we collected the corpus frequencies of the fillers of each of the 40 verb-slot pairs in the evaluation set. For this task, we needed a syntactically parsed corpus different from the treebanks we used as training set for the SP acquisition system. We decided to use two corpora that have been automatically parsed: the Perseus Digital Library and the Index Thomisticus; see section 3.2.1 for a description of these corpora. Two subsets of over 3 million tokens each were extracted and parsed with MST (McDonald et al., 2005) and MaltParser (Nivre and Scholz, 2004) respectively. ${ }^{35}$ From the parsed corpora, we extracted the set of fillers for each verb, together with their frequencies, provided that the fillers did not occur in the training set.

Maximum-frequency fillers vs. minimum-frequency fillers Now the question is to what extent this frequency distribution is reflected by the SP probability distribution assigned by our system. In order to test this, we first checked if higher frequencies corresponded to higher SP probabilities. In 27 cases out of 40, the mean of the probabilities assigned to the minimum-frequency fillers was lower than the probability assigned to the maximum-frequency fillers, i.e. there were 27 successes. A significance test showed that the successes were significantly more likely than the failures; see section B.9.1 for details.

Correlation test Subsequently, we performed a correlation test to quantify the association between the two distributions: the corpus frequency distribution and the SP probability distribution. In particular, we performed an ordinal test which considers the rankings assigned by the two distributions to the fillers extracted from the parsed corpora. The results are reported in section B.9.1 on page 223 and show that in a significantly higher proportion of cases ( 28 out of 40 ) there is no correlation in the rankings of the two distributions. This might be due to the size of the input data, which depends on the number of fillers extracted from the parsed corpora for each verb. This result shows that, although we are able to accurately detect the minimum and maximum-frequency fillers in most cases by simply judging from their SP probabilities, our accuracy is significantly lower when trying to assign the same ranking as the frequency distribution.

Fillers vs. "non-fillers" Finally, we compared the ability of our system to distinguish between plausible and implausible argument fillers. For each pair $(v, a)$ we calculated the mean probability $P_{a}(f \mid v)$ of all random fillers extracted from the parsed corpora for $(v, a)$ and compared it with the mean probability $P_{a}(n \mid v)$ of the random fillers extracted for the other verb-slot pairs. In this comparison 36 cases turned out to be successes, i.e. the probability of the actual filler was higher than the probability of the "non-fillers"; this is a significantly high proportion, providing a positive result. See section B.9.1 on page 226 for details.

[^45]
### 5.4.3 A distributional measure of semantic similarity

In the standard version of the algorithm the synset score and semantic score were chosen as measures of the similarity between verbs and nouns, respectively: cf. formula B.14. These scores rely on the structure and content of LWN, so that a high score means that the two words are highly related with respect to the WN-relations.

According to the syntax-semantics interface hypothesis (cf. section 2.2.2.3), two verbs sharing syntactic features such as SC patterns also share meaning components (Levin, 1993; McCarthy, 2001; Schulte im Walde, 2003). By combining this hypothesis with a distributional approach to semantics, we provided the algorithm for SP acquisition with an alternative measure of semantic similarity; this measure is calculated between words represented as vectors in vector space models based on the frequencies of their distribution in the corpora: see section B. 4 for an introduction to vector spaces.

In order to increase the frequency counts and get a better performance of the vector space model, we used the automatically parsed data described in section 5.4.2.1. The two corpora were then given as input to the database queries for the extraction of the SC lexicon, described in chapter 4. The new lexicon obtained this way were the input for the representations in the vector space model.

We represented each verb as a vector in a multi-dimensional space whose dimensions are all the possible SC structures observed in the lexicon; the coordinate of a verb vector in the $i^{\text {th }}$ dimension is the frequency with which the verb displays the $i^{\text {th }}$ syntactic pattern. Table 5.16 shows two example vectors for the verbs voco 'call' and nomino 'name'. Similarly, we represented the argument fillers occurring in the lexicon as vectors in the space defined by their occurrence with all the verb-slot pairs. Table 5.17 displays the two vectors for terra 'earth' and homo 'man'.

The cosine between two vectors is defined in section B.5.1: it is independent of the norm (size) of the vectors and ranges from 0 (when the two vectors are orthogonal) to 1 (when the vectors are aligned). The cosine distance between two vectors is defined as 1 minus the cosine between the vectors. We used the cosine distance between pairs of verbs and pairs of nouns in B. 13 as an alternative version of the synset score and the semantic score defined in section 5.3.2.1. For the verbs in table 5.16 the cosine is 0.59 , and hence their cosine distance is $1-0.59=0.41$. For the two nouns in table 5.17 , the cosine distance is 1 .

Central tendency We first compared the probabilities assigned by this version of the algorithm to the random fillers with the three measures of central tendency previously described. Of all 39 cases, 18 showed a probability of the random filler higher than the mean, 35 showed a probability higher than the median and 17 a probability higher than the value at the third quartile. The results from significance tests are provided in section B.9.2 and show that only the results for the median were positive (i.e. they had a significantly higher proportion of successes than failures). If we compare these results with those for the WN version of the algorithm we get: 18 vs. 28 (mean), 35 vs. 36 (median) and 17 vs. 24 (third quartile). In the WN case, two (mean and median) out of three measures showed a significantly higher proportion of successes, while the distributional version only has one positive result (median).

Random filler vs. remaining 39 random fillers We then compared the probability assigned by the system to each verb's random filler with the mean of the probabilities assigned to the 39 random fillers extracted for the other verbs. Our criterion for success was that the probability of the former is higher than that of the latter.

|  | $S_{1}$ | $S_{2}$ | $S_{3}$ | $S_{4}$ | $S_{5}$ | $S_{6}$ | $S_{7}$ | $S_{8}$ | $S_{9}$ | $S_{10}$ | $S_{11}$ | $S_{12}$ | $S_{13}$ | $S_{14}$ | $S_{15}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| voco | 1 | 1 | 5 | 1 | 0 | 0 | 5 | 0 | 1 | 2 | 0 | 1 | 1 | 10 | 3 |
| nomino | 0 | 0 | 3 | 1 | 3 | 1 | 4 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 |

Table 5.16: Vectors of the verbs voco 'call' and nomino 'name' in a multidimensional space defined by the SC structures with which these two
verbs occur in the two different 3-million-token parsed corpora. See table 5.18 for the description of the SC structures $S_{1}$, etc.

|  | $S_{1}$ | $S_{2}$ | $S_{3}$ | $S_{4}$ | $S_{5}$ | $S_{6}$ | $S_{7}$ | $S_{8}$ | $S_{9}$ | $S_{10}$ | $S_{11}$ | $S_{12}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| terra | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| caelum | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 1 |

Table 5.17: Vectors of the nouns terra 'earth' and caelum 'sky' in a multidimensional space defined by the first (verb, slot) pairs with which these nouns occur in the two different 3-million-token parsed corpora. See table 5.19 for the description of the structures $S_{1}$, etc.

| $S_{1}$ | A_(ad)Obj[acc],Obj[acc] |
| ---: | :--- |
| $S_{2}$ | A_(in)Obj[acc],Obj[acc],Sb |
| $S_{3}$ | A_Obj[acc] |
| $S_{4}$ | A_Obj[acc],Obj[dat],Sb |
| $S_{5}$ | A_Obj[acc],OComp[acc] |
| $S_{6}$ | A_Obj[acc],OComp[acc],Sb |
| $S_{7}$ | A_Obj[acc],Sb |
| $S_{8}$ | P_(a)Obj[abl] |
| $S_{9}$ | P_(a)Obj[abl],Sb |
| $S_{10}$ | P_(ad)Obj[acc],Sb[acc] |
| $S_{11}$ | P_(per)Obj[acc],Sb[nom] |
| $S_{12}$ | P_Obj[dat] |
| $S_{13}$ | P_Pnom |
| $S_{14}$ | P_Pnom,Sb |
| $S_{15}$ | P_Sb |

Table 5.18: Description of the SC structures in table 5.16.

In a significantly higher proportion of cases ( 25 out of 39 ) we got a success. This result is slightly (but not significantly) higher than 24 , the result of the same evaluation measure for the WN version of the algorithm.

### 5.4.3.1 Fillers from parsed corpora

We collected the corpus frequencies of the fillers of each of the 40 verb-slot pairs in the evaluation set from the two afore-mentioned parsed corpora.

Maximum-frequency fillers vs. minimum-frequency fillers In 18 cases out of 40, the mean of the probabilities assigned to the minimum-frequency fillers was lower than the probability assigned to the maximum-frequency fillers, i. e. there were 18 successes. A significance test showed that the successes were significantly less likely than the failures, unlike the results of the same test on the WN version of the algorithm. See section B.9.1 for details.

Correlation test We performed an ordinal correlation test between the two distributions, corpus frequency distribution and SP probability distribution. The results are reported on page 226; they show that in a significantly higher proportion of cases ( 24 out of 40 ) there is no correlation in the rankings of the two distributions.

Fillers vs. "non-fillers" We then checked if the fillers received higher probabilities than "non-fillers". In 32 cases out of 39, i.e. in a significantly higher proportion of cases, this was the case; see page 226 for details. This positive result corresponds to a similar positive result for the WN algorithm.

Overall, we can say that the WN version performs better. Further work on testing different distributional measures of similarity and/or data parsed with a different parser may show an opposite outcome.

| $S_{1}$ | abeo+A_Obj[acc] |
| ---: | :--- |
| $S_{2}$ | abeo+A_Sb[nom] |
| $S_{3}$ | adiuvo+A_Sb |
| $S_{4}$ | admiror P_Sb |
| $S_{5}$ | aduro+A_Obj[acc] |
| $S_{6}$ | aperio+A_Sb |
| $S_{7}$ | ascendo+A_(de)Obj[abl] |
| $S_{8}$ | ascendo+A_(in)Obj[acc] |
| $S_{9}$ | cado+A_(de)Obj[abl] |
| $S_{10}$ | cado+(in)Obj[acc] |
| $S_{11}$ | cado+A_(super2)Obj[acc] |
| $S_{12}$ | claudo+A_Obj[acc] |

Table 5.19: Description of the structures in table 5.17.

### 5.5 Conclusions and future work

We presented how we adapted the method by Alishahi (2008) to our Latin data. The peculiarities of the data regard the goals of the analysis, the size of the corpora and the resources used.

As stated in the introduction chapter on page 6, our first goal is lexicographic, i.e. to acquire SPs that can be included in a lexical resource for Latin verbs such as the SC lexicon. As a consequence of this, we focussed on finding a way to optimise the end result and its integration with the input syntactic information from the SC lexicon.

Our second goal is methodological and addresses both Latin linguists and computational linguists, as we said on page 6 . We customised Alishahi's model to the small corpora available to us and to the low coverage of LWN, showing that research on computational semantics can benefit from this work. For example, the generalization from seen instances to SPs was made easier by expanding the set of occurrences used by the system for calculating the SPs. Unlike Alishahi's approach, we decided to consider the occurrences of synonyms of the target verb as well as the occurrences of that verb. This could work as well for less-resourced languages other than Latin. In addition, we introduced morphological attributes of verbal arguments among the syntactic features of the frames; moreover, we accounted for partial matches between these features, given the peculiarities of Latin syntax and its free word order. This tailored the algorithm to Latin; however, similar changes can apply to other (ancient) languages.

The nature of Latin as a dead language affected the evaluation as well, which was performed on corpus data rather than on native speakers' intuition. We chose three measures of central tendency to detect the cases where a random filler manually extracted from a large corpus occurring in the test corpus fell into the high-probability region of the SP distribution given by our model. In addition, we counted how many times a random filler for a verb-slot pair was assigned a higher probability than the mean of the random fillers for the remaining pairs. Moreover, we exploited the frequencies from a large automatically parsed test corpus in order to test whether our system was able to distinguish between more plausible fillers and less plausible ones, and even their internal ranking. Finally, we compared the probabilities of the actual fillers of a verb-slot pair in the parsed corpus with those of the fillers for the other pairs.

In the evaluation phase we had the chance to show an interesting comparison between
the original, knowledge-based version of the algorithm, which uses LWN-based measures of semantic similarity, and a knowledge-free version of the algorithm, which uses a distributional measure based on the cosine distance. The results show that the original version succeded in 5 evaluation tests out of 7 , while the distributional one failed in 4 . The test based on one measure of central tendency (third quartile) as well as an ordinal correlation test on the filler rankings gave negative results in both cases.

Overall, the WN algorithm performs better, which might be due to the fact that the parsed data are noisy, and this affects the filler extraction step, on which the similarity measure relies. The noise can be partially removed by more sophisticated parsers, which take into account features specific to Latin, such as the parser described in Passarotti and Dell'Orletta (2010). In addition, distributional measures other than cosine should be tested, to see if the cosine distance was actually the best option. A list of alternative distances is provided in Erk (2007), for example.

Our motivation relied on the clustering step as a way to help the generalization from seen fillers to SP classes. After implementing the algorithm, we found out that the clustering step does not affect the performance of the system as measured by the chosen evaluation tests. This could change with a larger dataset and by changing the smoothing factors to get larger constructions.

The results are overall encouraging, showing a good calibration and a good discrimination of the system. However, the performance can still be improved: the current best version, the WN-based one, assigns a SP probability to an LWN node based on its LWN similarity with the fillers of verbs similar to the target verb; consequently, the results depend both on the occurrences in the input corpus and on the structure of LWN. Varying the input corpus affects the set of seen instances but not the way they contribute to the probability assigned to similar nouns. Instead, a combination of the LWN measure and a distributional measure could account for the distinctive features of the input corpus in the similarity measure, too. This may allow for more direct comparisons between results from corpora with different features for genre, era, authors, etc.

## Chapter 6

## Introduction to Latin preverbs

### 6.1 Introduction

The present chapter gives an introduction to the linguistic phenomenon which will be analysed in chapter 7: the argument structure of Latin preverbed verbs (PVs). Therefore the exposition is intended to serve as a background to the empirical investigation in chapter 7. Since the investigation takes a theoretically informed data-driven approach, in accordance with the overall methodological aim of the thesis (cf. section 1.2 ); the background discussion will necessarily focus on issues that are directly related to this approach rather than going into the full details of the theoretical views.

Section 6.2 gives a first definition of preverbs. Section 6.3 is a survey of the state of the art on Latin preverbs and summarises the different theoretical approaches to this topic. Finally, section 6.4 focusses on the properties of the argument structure of PVs in terms of relatum and locatum and prepares the ground for the analyses illustrated in chapter 7 .

### 6.2 A first definition of preverbs

Morphologically, we define preverbs in Latin as prefixes of verbal stems. For example, the preverb $a b$ - 'from' is prefixed to the verb eo 'go' in the PV abeo 'go from, depart'.

In Booij and van Kemenade (2003) preverbs are defined as "morphemes that appear in front of a verb, and which form a close semantic unit with that verb". This definition makes an explicit link between the morphological origin of preverbs and their semantic properties with respect to the verb they are prefixed to. We will come back to this link in connection with the verb's argument structure in section 6.4.

In this chapter we will study the realisation of the spatial arguments of Latin PVs by focussing on preverbs acting as bound prefixes; however, from a diachronic point of view, we are also interested in the connection between these preverbs and free morphemes such as adpositions, to the extent that this connection is reflected in the argument realisation of PVs, and more specifically in the morpho-syntactic realisation of the motion events evoked by the PVs.

### 6.3 Previous work on preverbs

This section gives an overview of the state of the art on preverbs. Section 6.3 .1 briefly reviews some of the main contributions to the study of Latin preverbs in general, while section 6.3.2 goes into the discussion on the origin and development of Latin preverbs. Finally, section 6.3.3 frames preverbs within a theory of motion events and spatial expressions.

### 6.3.1 Overview of the literature

Preverbs have been treated within several linguistic subfields, referring to various theoretical frameworks and offering different definitions.

The amount of studies on this topic is very large and a full account of them falls outside the scope of this work. For this reason, in the present section we focus on Latin preverbs and briefly sketch some of these approaches: historical linguistic studies (section 6.3.1.1) and typological studies (section 6.3.1.2). Finally, we state the scope of our research (section 6.3.1.3). The aim of this overview is to illustrate the complex nature of the phenomenon of preverbs and emphasize its possible ramifications.

### 6.3.1.1 Historical linguistic studies

In this section we briefly survey some of the main contributions to the study of Latin preverbs in the field of historical linguistics. The survey includes diachronic studies in Indo-European and Romance linguistics, as well as syncronic studies in Latin linguistics, which will be further discussed in sections 6.3.2.1 and 6.3.2.2.

Indo-European linguistics Preverbs have often been treated together with adpositions, due to their formal similarity and their shared historical evolution. A common origin of preverbs and adpositions from old Indo-European free adverbial morphemes has been claimed by several scholars (Benveniste, 1949; Lehmann, 1983; Echarte Cossío, 1998; López Moreda, 1998; Mellado Rodríguez, 2001; Cuzzolin et al., 2006). In particular, it has been hypothesized that these free elements subsequently developed into bound morphemes, thus becoming inseparable from the verbs. The discussion in the above works involves theories on the grammaticalization of adpositions into preverbs as well as investigations on the evolution of word order, case system and prepositions in the ancient Indo-European languages (Cuzzolin, 1995; Fruyt, 1998; Luraghi, 2010b). We will cover these issues in section 6.3.2.1 and in section 6.3 .3 we will see that from a theoretical point of view the features shared by preverbs and adpositions make their categorisation particularly difficult.

Romance linguistics Latin preverbs have also been studied in a diachronic perspective within Romance linguistics. In particular, the lexicalization of Latin PVs in Italian has been connected with the origin of Italian phrasal verbs (Masini, 2005; Iacobini, 2005; Simone, 2008). The process of lexicalization of preverbs has also been interpreted in a functionalist perspective by Imbert (2008). In section 6.3.2.2 we will go into this question in more detail.

Phonology and morphology Different views have been proposed to classify the phenomenon of preverbation in relation with other morphological phenomena. Oniga
(2005) contrasts prefixation and preverbation with composition, giving a set of phonomorphological and semantic features that distinguish these two phenomena. On the other hand, Heslin (1987)'s phonological analysis constrasts prefixation (and preverbation) with derivation, viewing prefixation as "a form of compounding in Latin".

Actionality Preverbs have been seen as Aktionsart markers in various ancient and modern languages. For what concerns Latin, atelic events represented by non-prefixed verbs have been contrasted with telic events in the corresponding PVs: cf. facio, 'make', 'do' vs. perficio, 'achieve' (Romagno, 2004). Moreover, preverbs have been claimed to describe different stages of a verbal process: the allative/ablative opposition between preverbs ( $a d$ - and in- 'to' vs. $a b$-, ex- and $d e$ - from') has been viewed as an aspectual opposition between ingressive and resultative (García-Hernández, 2005). From this perspective, an allative preverb such as ob- can mark an ingressive aspect in a verb like oppugno 'assail', and an ablative preverb such as ex- can modify a verb like molior 'labour' towards a resultative aspect in emolior 'accomplish'. This approach will not be dealt with in our analysis.

Synchronic studies Synchronic studies in Latin linguistics have focussed on the system of Latin prepositions and cases (Benveniste, 1949; Echarte Cossío, 1998; Carvalho, 1998) and on the semantics of prepositions, preverbs and PVs (García-Hernández, 2005). Furthermore, several studies in Latin philology have been devoted to single preverbs or to groups of PVs, in order to investigate their semantics in a synchronic or diachronic perspective (Van Laer, 1998, 2005; Brachet, 2005; Calboli, 2005; Gaide, 2005; Molinari, 2005; Longrée, 2005; Bortolussi, 2005; Moussy, 2005). In particular, the argument structure of PVs has been analysed (Bortolussi, 2005) and contrasted with the argument structure of simple verbs (Lehmann, 1983; Touratier, 2001; Longrée, 2005). Section 6.4 will develop the topic of argument structure, which is the object of our study.

### 6.3.1.2 Typological studies

This section reports on selected typological studies on preverbs.

Cross-linguistic studies Preverbs are crosslinguistically (but not necessarily universally) attested and some attempts have been made to describe common properties of preverbs in different languages. Booij and van Kemenade (2003) give examples from some Indo-European, Finno-Ugric, Georgian and Caucasian languages, while Rousseau (1995) focusses on ancient and modern European languages and Heine et al. (1991) deal with specific African languages.

As we said, our interest will be in Latin and to some extent in certain Indo-European languages - such as Ancient Greek and some Romance languages - that help understand the evolution of Latin preverbs, although the present study is not typological in nature.

Motion events Spatial preverbs have been discussed within the typology of motion events (Talmy, 2000; Meini, 2009; Spreafico, 2009) and particularly in the distinction between satellite preverbs and relational preverbs (Craig and Hale, 1988; Imbert, 2008). This question is dealt with in section 6.3.3, in connection with grammaticalization and lexicalization processes.

### 6.3.1.3 Our focus

In chapter 7 we carry out a syntactically-based quantitative analysis on the arguments of PVs as they are encoded in the annotation of the Latin treebanks we use and as they are provided by the subcategorization (SC) lexicon described in chapter 4. As pointed out in section 2.2, there is considerable debate over the question of argument structure, although a full treatment of the theoretical status of argument structure falls outside the scope of the present study. Instead, we will take a data-driven approach by using our lexicon. There are primarely two advantages that follow from this methodological decision. First, the annotation was not done specifically with preverbs in mind, which reduces potential biases; second, using an existing annotation leads to replicable and more transparent results.

In our analysis, the semantic features of PVs will be investigated with respect to their interaction with the syntactic realisation of the arguments. The theoretical background to our diachronic analysis is provided by theories on the origin and evolution of preverbs (section 6.3.2), as well as grammaticalization theory (section 6.3.2.2) and typology of space (section 6.3.3). This eclectic use of theory is motivated by the datadriven approach discussed above, where the theories serve as a context for the empirical investigation.

### 6.3.2 Origin and development of Latin preverbs

In this section we review some of the theories proposed to explain the origin and development of Latin preverbs within Indo-European linguistics (section 6.3.2.1). Finally, we touch upon studies on grammaticalization and lexicalization of preverbs in section 6.3.2.2. For a discussion of these issues see also Meini and McGillivray (2010).

### 6.3.2.1 Indo-European particles and Latin preverbs

A hypothesis in Indo-European linguistics that goes back to Saussure (1968, pp. 216-217) and was drawn on by Meillet and Vendryes (1963, pp. 199-200, 573-578) associates preverbs to adpositions ${ }^{1}$ through their common origin from a class of undifferentiated proto-Indoeuropean adverbial particles: the so-called ADV/ADP/PREV category (Cuzzolin et al., 2006). ${ }^{2}$ According to this hypothesis, such original adverbial elements were characterized by a rather free status within the sentence both as to their linear order and their syntactic functions; ${ }^{3}$ in addition, they carried a spatial or temporal notion and specified the generic role of cases.

In Vedic such particles are often separated from the verbs (MacDonell, 1953, p. 285), while this is not the case in Classical Sanskrit (Whitney, 1989, pp. 414 ff .). Already in Rigveda Vedic preverbs and verbs seem to form a morpho-syntactic unit, especially for what concerns the stress, which is on preverbs in main clauses and on verbs in sobordinate clauses.

Homeric Greek too displays various examples of the flexible status of such elements. ${ }^{4}$ For instance, sentence (1) allows for two different parsings, depending on whether we

[^46]interpret $e k$ 'from' as attached to the noun nèòs (adpositional interpretation) or to the verb bê (preverbal interpretation).

| (1) ek dè Chrysḕs nēò | bê |
| :--- | :--- |
| from and Chryseis:NOM.M.SG ship:GEN.F.SG go:IND.AOR.3SG seafaring:GEN.F.SG |  |
|  | 'Chryseis went out of the sea-crosser ship'/ 'Chryseis disembarked from the sea-crosser ship". |

A first phase of development of preverbs and adpositions from independent constituents is attested in Latin as well, but evidence is limited to a few archaic texts. As an example, we will consider the following passage by the grammarian Festus quoted in Cuzzolin (1995, p. 130):
(2) Sub vos placo, in precibus fere cum dicitur, significat id, quod supplico, ut in legibus: transque dato et endoque plorato.
'When we say, mostly in prayers, sub vos placo, it means the same as supplico, like transque dato and endoque plorato in the laws.'

In this passage, while talking about the language of early prayers, Festus sets a semantic equivalence between the classical form supplico 'implore, beseech' and its $a$ posteriori decomposition into the elements sub 'under' (adverb in first position) and placo 'reconcile' (verb); in this decomposition, the enclitic pronoun is inserted in the second position, according to Wackernagel's law (Wackernagel, 1892).

As to early Latin, in a study on intro- Brachet (2005) shows that in Plautus and Terence intro- occurs both as a preverb and as a free adverbial element. In spite of such few examples, in Latin we typically find a clearly delimited situation with prepositions on the one hand, and preverbs as bound prefixes on the other. We will see that, although this distinction is relatively clear at the morphological level, it is not always as clear from the syntactic and semantic point of view.

The situation for Latin has been explained diachronically through a combination of processes referring to word order change and the erosion of the case system (Cuzzolin, 1995; Fruyt, 1998; Oniga, 2005; Luraghi, 2010b), although the individual stages of this evolution are not uncontroversial. These scholars have suggested that, as the order of words became more fixed, the original adverbial elements could either be morphologically fused with the following verbal forms (univerbation) thus developing into preverbs, or they could syntactically govern the case of the preceding nouns, thus generating postpositions. In the latter case, the semantic constituents made of adverbial particles and inflected nouns frequently occurring together might have gradually acquired the status of syntactic constituents, as the syntactic connection between prepositions and nouns' inflection became more and more constrained (Luraghi, 1989).

This double development represents two parsings of the same original structure $\mathrm{N}+\mathrm{Adv}+\mathrm{V}$, where 'Adv' stands for an adverbial element, ' N ' for a noun and ' V ' for a verb: the $\mathrm{N}+[\mathrm{Adv}+\mathrm{V}]$ analysis would explain the rise of preverbs (3), while $[\mathrm{N}+\mathrm{Adv}]+\mathrm{V}$ would account for the genesis of postpositions (4). ${ }^{5}$
(3) flumen adeo
river:ACC.N.SG go to:IND.PRS.1SG
(4) *flumen ad eo
river:ACC.N.SG to go:IND.PRS.1SG

[^47]In connection with the hypothesized typological shift from OV to VO order, some postpositions (flumen ad) are said to have later turned into prepositions (ad flumen) (Oniga, 2005, p. 223).

The linguistic changes affecting word order, preverbs and adpositions must have included the overlapping transitional stages acknowledged by Heine et al. (1991) and Magni (2008). This evolution was in fact complex, as shown by the high number of competing explanations proposed for it.

Regarding preverbs and word order, Cuzzolin et al. (2006) underline how the development of preverbs is not contradicted by a dominant VO order, but "it is rather the syntactic position that assigns the functional value of ADV or ADP", referring to the categories of adverbs and adpositions respectively. Magni (2008) proposes a different view on Latin word order, providing a detailed discussion of this complex and multilevel evolution, and accounting for the different factors affecting it. She concludes that the role played by syntactic and semantic parameters affecting word order is connected with the absence of a unique "holistic" model for the evolution of Latin word order.

The position according to which adpositions are related to word order changes is criticized by Baldi (1979). Baldi proposes that the unmarked position of the original free adverbial particles in Indo-European was preverbal. From these particles, preverbs developed as bound prefixes; prepositions thus originated from the fact that these preverbs could govern the case of the NPs accompanying the verb.

Luraghi (2010b) includes these phenomena in a more general discussion on nonconfigurational constructions in old Indo-European languages. Non-configurational languages are said not to have a hierarchical structure at the Phrase Structure level, but only at the Lexical Structure; free word order and the use of syntactically discontinuous constituents, among others, are considered typical non-configurational properties. In this view, adpositional phrases can be considered configurational constructions.

These complex phenomena - the rise of fixed word order, emergence of analytical structures and weakening of the case system - would deserve a much longer discussion which falls outside the scope of this work. The brief overview we gave serves as a general introduction to how different linguistic theories have dealt with the phenomena involving preverbs.

### 6.3.2.2 Grammaticalization

Grammaticalization theory offers a theoretical framework for describing the evolution of Latin preverbs and adpositions, as discussed in Cuzzolin (1995); Fruyt (1998) and Luraghi (2010b).

The term grammaticalization has been used to refer to different models of language change, ranging from Meillet (1921) to the more recent developments and theoretizations (Traugott, 2006). A commonly used definition of grammaticalization is given in Kuryłowicz (1975, p. 54).

Under the diachronic aspect, grammaticalization is a process which turns lexemes into grammatical formatives and makes grammatical formatives still more grammatical.

Although the phenomenon of grammaticalization is widely accepted, its explanatory value is disputed: see Campbell (2001) for a critical discussion. Our study on Latin preverbs reported on in chapter 7 aims at establishing the diachronic patterns for PVs

# preverbs <br> spatio-temporal particles $\rightarrow$ adverbial elements <br> adpositions 

Table 6.1: Grammaticalization chain for the origin of preverbs in Latin.
attested in the corpus evidence we used; it does not concern itself with a causal explanation for the grammatical changes PVs underwent. In this perspective, we refer to grammaticalization theory as a descriptive device.

Grammaticalization studies have been devoted to the analysis of preverbs and adpositions: cf. e.g. Heine et al. (1991) for studies of some African languages. In the case of Latin preverbs, an origin from lexical particles with local or temporal meaning has been hypothesized for the original Indo-European adverbial elements treated in Cuzzolin (1995) and Cuzzolin et al. (2006), as we saw in section 6.3.2.1. A grammaticalization account of the origin of preverbs and adpositions from these adverbial elements would place these elements at the beginning of a grammaticalization chain, for which table 6.1 gives a schematic illustration. It should be kept in mind here that grammaticalization is said to consist of gradual overlapping stages (Lichtenberk, 1991) which are not reproduced in table 6.1.

From the undifferentiated class of adverbial elements in Indo-European the chain would continue by showing the distinct evolution of preverbs and adpositions towards "even more grammatical" items. Such differentiation preverbs and adpositions underwent is related to the crucial possibility of attaching these elements to VPs or NPs respectively, as described in section 6.3.2.1. This emphasis on the role played by constructions ${ }^{6}$ and context is the main idea in the following alternative definition of 'grammaticalization' (Hopper and Traugott, 1993, p. xv) as:
the process whereby lexemes and constructions come in certain linguistic contexts to serve grammatical functions.

Under this perspective, the possibility that preverbs can govern the case of the verb's argument can be seen as an aspect of their grammaticalization manifested at the syntactic level, as we will see in section 6.4.

Further developments Here we will focus on further developments which have been hypothesized for Latin preverbs and PVs. The evolution of Latin preverbs seems to conform to a case of univerbation involving fusion of morphemes and loss of semantic compositionality. This process has elements of both grammaticalization and lexicalization.

As anticipated in section 6.3.2.1, after the univerbation of preverbs with simple verbs, in some occasions preverbs underwent a phonological fusion: cf. e. g. affero 'bring to' <ad-fero. In addition, the Latin PVs are generally said to have lost their semantic relation with the original verb so that their etymological meaning tended to weaken (cf. Cuzzolin et al. 2006, among others).

[^48]This process led some preverbs to take a 'more grammatical' role, distinct from that of prepositions, such as intensive, attenuative, iterative, or aspectual (Meillet and Vendryes 1963, pp. 301-302, López Moreda 1998). Cuzzolin (1995) proposes an evolution of some preverbs to aspectual markers with no lexical content: cf. com-edo (lit. 'eat with') which means 'eat up', not 'eat with someone'. This semantic phenomenon can be identified as an instance of semantic bleaching, defined in Matisoff (1991, p. 384) as "the partial effacement of a morpheme's semantic features, the stripping away of some of its precise content so it can be used in an abstracter, grammatical-hardware-like way". Semantic bleaching is usually considered as one of the features of grammaticalization processes because it implies a replacement of a "lexical content meaning" with a "new, more abstract, grammatical meaning" (Brinton and Traugott, 2005, p. 29), as in the typical example of the English grammaticalized form going to from the motion sense to the future marker. Here we will consider the term 'semantic bleaching' as a metaphor and a heuristic device, since its explanatory status is debated (Jenset, 2008).

Besides instances of semantic bleaching and thus grammaticalization, we can observe another group of phenomena affecting Latin preverbs. As a matter of fact, some Latin preverbs did not acquire an aspectual sense. For these preverbs a development towards a non-compositional meaning of the PVs has still been claimed (López Moreda, 1998). Consider, for example, ob-ligo, whose etymological sense 'bind around' is rarer than the sense 'oblige, make liable', where the prefix does not hold an explicit semantic content. Likewise, permitto originally meant 'send through', but by the first century B.C. it acquired the meaning 'allow', cf. French permettre and Italian permettere (Vincent, 1999). As will be demonstrated below, the development of these preverbs displays some features of lexicalization.

According to the controversial principle of unidirectionality in grammaticalization theory, "grammaticalization is a unidirectional process; that is, it leads from less grammatical to more grammatical forms and constructions" (Heine and Kuteva, 2002, p. 4). Under this perspective the opposite of grammaticalization is called lexicalization, ${ }^{7}$ and diachronically proceeds from a grammatical word or a phrase to a lexical word. ${ }^{8}$ Partially overlapping with both these phenomena is degrammaticalization, sometimes considered as a wider phenomenon which includes lexicalization: a typical example is the noun ism 'ideology' from the suffix -ism (Ramat, 1992; Lazzeroni, 1998).

Lexicalization has been widely debated: for an overview of theories on lexicalization and its integration with grammaticalization see Brinton and Traugott (2005) and Fischer (2007). Here we will outline some lexicalization phenomena regarding certain Latin preverbs and their further evolution in neo-Latin languages to the extent that this is relevant for the corpus-based study of PVs' argument structure reported on in chapter 7.

Studies on the origin of verbal prefixes in Romance languages have claimed that Latin preverbs underwent a lexicalization process within the PV they were prefixed to, which has been connected with a gradual loss of transparency of Latin preverbs inside the PV (Tekavčić 1972, §948.3-1345, Salvi and Vanelli 1992, p. 206, Crocco Galèas and Iacobini 1992, p. 172, Vicario 1997, p. 129, Haverling 2000, pp. 458-460, Dufresne et al. 2003, p. 40). This phenomenon does not affect all preverbs in the same way and it is said to be more pronounced for some preverbs - such as ad- 'to', in- 'in, to' - and less for some other ones - such as de- 'from'. Allen (1981, p. 84) connects this with the phonological combination between preverb and verb stem. For example, consonantal

[^49]assimilation and loss of final consonants in prepositions caused verbs prefixed with the preverbs $a b$ - 'from' and $a d$ - 'to' to converge into the same class.

Specific interwoven phenomena have been detected that attest a lexicalization process which has been related with the gradual loss of Latin case system and the subsequent increase of analytic forms involving prepositions in neo-Latin languages (Iacobini and Masini, 2007).

First, the productivity of some Latin locative preverbs contrasts with their low productivity or non-productivity in Romance languages (Tekavčićc 1972, §948.3-1345, Dardano 1978, pp. 123-137, Allen 1981, Serianni 1989, pp. 660-661, Lüdtke 1996, Vicario 1997, p. 129, Iacobini 2005). ${ }^{9}$ For instance, certain Latin preverbs ( ad - vs. in-, $d e$ - vs. ex-) indicate different spatial directions: the difference between $a d$ - and in- can be explained through the opposition between approaching and entering a limit; on the other hand, $a b-$-, as opposed to $a d$-, expresses a departure, while ex- expresses an exit from a place (García-Hernández, 2004). ${ }^{10}$ In the Romance languages this functional distinction is mostly not present any more: cf. French attendre 'wait' vs. entendre 'hear' (Iacobini, 2005). ${ }^{11}$

Second, Durante (1981, p. 66) connects the loss of productivity of some Latin preverbs with the switch to SVO word order that caused the replacement by analytic forms: cf. subire>ire susu(m) 'go from below upwards' (Masini, 2005, p. 161). The use of such postverbal particles instead of preverbs is connected with the later rise of phrasal verbs in some Italian dialects (Simone, 1997; Vicario, 1997; Masini, 2005): ${ }^{12}$ cf. section 6.3.3.2.

In conclusion, it seems reasonable to hypothesize a connection between the gradual lexicalization of Latin preverbs - already attested in the Classical era - and an increasing use of analytic forms such as prepositions and adverbial elements to express directional meanings (Touratier, 1994, pp. 383-384), which finally led, among other things, to the emergence of Italian phrasal verbs (Masini, 2005). As we will see in section 6.4, our interest is in the extent to which this evolution affected the argument structure of Latin PVs, with particular reference to the use of PPs as spatial complements.

### 6.3.3 Spatial preverbs and motion events

In section 6.3 .2 we stressed the fuzziness of the category of the adverbial elements which originated preverbs and adpositions, according to theories in Indo-European studies. In sections 6.3.3.1 and 6.3 .3 .2 we will see how this fuzziness also applies to the category of Latin preverbs, once we define them according to Talmy's typology of space (Talmy, 2000).

[^50]
### 6.3.3.1 Motion events

Talmy's theorisation of spatial schemas focusses on closed-class forms and the relation between the semantic and the morphological level in the expression of motion events (Talmy, 2000, p. 25):

The basic Motion event consists of one object (the Figure) moving or located with respect to another object (the reference object or Ground). It is analysed as having four components: besides Figure and Ground, there are Path and Motion. The Path (with a capital P) is the path followed or site occupied by the Figure object with respect to the Ground object. The component of Motion (with capital M) refers to the presence per se of motion or locatedness in the event. [...] In addition to these internal components, a Motion event can be associated with an external Co-event that most often bears the relation of Manner or of Cause to it.
In addition, locations are defined as those space events where a movement is not realised. Hence, the spatial information is not always concentrated in one element only, but can be distributed along the different elements of a construction: cf. the so-called distributed spatial semantics (Sinha and Kuteva, 1995).

In the following sentences we distinguished between a Motion component expressing Location (5) and one expressing Translational Motion (6).
(5) The boy is in the room.

Figure Location Path Ground
(6) The boy is going into the room.

Figure Translational Motion Path Ground
(7) shows an example of a Co-event (Manner) expressed in the verb together with the translational motion, which makes it into an example of conflation.

```
The boy runs out of the room.
Figure Translational Motion+Manner Path Ground
```

The Path expressions can be of three types: FROM paths, TO paths and VIA paths. In particular, FROM paths refer to the Source from which the Figure moves, while TO paths refer to its Goal.

### 6.3.3.2 Satellites and adpositions

In sentences (5), (6) and (7) the Path component is encoded outside the verbal root in the prepositions in, into and out of, which are the heads of the PPs in the room, into the room and out of the room.

This is not the only possibility. Sentence (8) shows an example of a phrasal verb where the particle out relates to the verb run and contains the Path of the motion event.

The boy ran out.
Figure Translational Motion+Manner Path
Talmy (1991, p. 486) defines particles like out in the English phrasal verb run out as satellites, i.e.
a grammatical category of any constituent other than nominal complement that is in sister relation to the verb root. The satellite, which can be either a bound affix or a free word, is thus intended to encompass all of the following grammatical forms, which traditionally have been largely treated independently of each other: English particles, German separable and inseparable verb prefixes, Latin or Russian verb prefixes, Chinese verb complements, Lahu non head versatile verbs, Caddo incorporated nouns, and Atsugewi polysynthetic affixes round the verb root.
satellites complement the content of the verb from a lexical and semantic point of view, and - unlike adpositions - do not govern an NP. In spite of this categorisation, particles in phrasal verbs have been recognised to bear an "ambiguous categorical status" between prepositions and adverbs, and the resulting multi-word expressions may assume a metaphorical non-compositional meaning (Masini, 2005, p. 155). satellites, in fact, may lexicalize ${ }^{13}$ with the verb and lose their spatial meaning, as in (9).
(9) We ran out of sugar.

Talmy's typology Talmy (2000) presents a semantic typology based on the different ways languages morphologically encode the Path component of motion events through so-called lexicalization patterns. According to his typology, most languages belong to one of the two following groups:

- verb-framed languages: the Path component is in the verb root;
- satellite-framed languages: the Path component is in the satellite and/or preposition.

The first category of languages includes Romance languages as well as, Semitic, Japanese, Tamil and Polynesian languages. The second category includes Germanic, Finno-Ugric, and Chinese languages, among others. The difference between verb-framed and satellite-framed behaviour is exemplified by sentences (10) and (11).
(10) Il ragazzo corse fuori.
'The boy ran out'.
Il ragazzo uscì (correndo/di corsa).
'The boy exited (running)'.
In (10) the Path component is placed in the particles out and fuori, while Motion and Manner are encoded at the same time in the verbs run and correre (conflation); in (11) Motion and Path are fused into the verbs uscire and exit.

Several scholars have shown that Talmy's categorisation is too strict: a language can perform more than one strategy for encoding the Path component (Wälchli, 2000; Slobin, 2004). Both Italian and English, for example, allow the two encodings:

- satellite-framed: verbi sintagmatici in Italian (Rohlfs, 1970; Simone, 1997; Masini, 2005; Simone, 2008) ${ }^{14}$ and phrasal verbs in English (cf., for example, Dikken 1995).

[^51]An English example is go out; Italian examples are andare fuori 'go out', buttare via, which means 'throw away' but also 'waste'.

- verb-framed: Italian discendere and uscire, and English descend and exit.

Classifying Latin A strict interpretation of Talmy's classification implies that Latin belongs to the satellite-framed class, due to its category of preverbs. However, Romance languages are generally said to be verb-framed languages because they mostly encode Path inside the verb: cf. for example French sortir, Italian uscire 'exit'. This difference has been addressed in a number of studies proposing a more fine-grained classification, cf. e. g. Schwarze (1985) and Simone (1996).

Here we will concentrate on those Romance descentents of Latin PVs which have still kept a spatial meaning, although the role of the preverb may not be recognisable as a satellite any longer.

In her study on Late Latin PVs, Stolova (2008) states that the Romance verbs entrar (Spanish), entrer (French), entrare (Italian), entrar (Portuguese), a intra (Rumanian), entrar (Catalan), derived from Latin intrare 'enter, go in' (with preverb intro- 'inside') were etymologically compounds but perceived as simple forms. ${ }^{15}$ The same argument pertains to the Latin verbs ascendo 'ascend, go up', descendo 'descend, go down', exeo 'exit, go out' and pervenio 'arrive'. These verbs share the characteristic that their Latin preverbs "became non-transparent" and they survived in the Romance languages just as simple verbs like venio 'come' did.

In other words, Latin preverbs are defined as satellites by Talmy, but some Romance verbs from Latin PVs do not compositionally show the presence of the same preverbs as satellites. Therefore, we can hypothesize a development that turned Latin preverbs functioning as satellites into some Romance verbal prefixes that still encode Path but do not hold the satellite function any more. All these facts help frame the typological shift from Latin to Romance as a lexicalization process (see section 6.3.2 on page 124). In section 6.4 we will examine how this lexicalization can be partially identified by some properties of the argument structure of Latin PVs and which differences among Latin preverbs can help understand their problematic definition as satellites.

### 6.4 Argument structure of Latin preverbed verbs

In the present section we will survey some of the phenomena related to the argument structure of Latin PVs. In particular, we will concentrate on verbs of motion and verbs of rest, and on the realisation of their spatial preverbs' relata, i. e. the morpho-syntactic expresssions of the Path components corresponding to the preverbs. There has been a discussion in linguistics as to consider the status of Path expressions within the verbs' argument structure. In Jackendoff (1990)'s analysis, Path is not considered as a part of the argument structure of Manner of Motion verbs. Nevertheless, other verbs like buy, sell, give do require the Path component in their semantics and syntax. Since these specific Path expressions are annotated as verb arguments in the treebanks we use, we will consider the preverbs' relata as part of the argument structure of Latin verbs of motion and verbs of rest.

Latin PVs exhibit a variety of constructions in the realisation of their relata; these constructions are described synchronically in section 6.4.1; section 6.4.2 formulates some diachronic hypotheses which will be statistically tested on corpus data in chapter 7 .

[^52]
### 6.4.1 Synchronic overview

The following example illustrates the syntactic construction of the verb adeo 'go to', 'reach' with an accusative expressing the spatial complement (12), contrasted with the intransitivity of the simple verb eo 'go' occurring with a spatial complement introduced by ad (13):
(12) patriosque adit inpiger ortus (Ov., Met., I,
and fatherly:ACC.M.PL go to:IND.PRS.3SG quick:NOM.M.SG sunrise:ACC.M.PL 779)
'he rushed to his father's orient'
(13) ibat res ad summam nauseam (Petron., Sat, 78)
go:IND.IMPF.3SG thing:NOM.F.SG to highest:ACC.F.SG nausea:ACC.F.SG
'The thing was becoming extremely sickening'

In (12) the preverb corresponds to a preposition governing the accusative case (ad). In such cases, the valency of the PV seems to be somehow affected by the valency of the preverb, which causes an apparent "transitivisation" of the verb. This phenomenon has been explained from different points of view. For example, some Latin grammars use the notion of a transitive verb "evoked" by the PV (Ernout and Thomas, 1972, p. 21). In a generative framework, Oniga (2005) explains it as a consequence of the trace left by the preposition, which governs the accusative case. On the other hand, Romagno (2003) views the change of syntactic valency as an "epiphenomenon" of the change in the actionality of the verb, which operates on the direct internal argument of the verb.

As a matter of fact, a relation can be established between the argument structure of a PV and the argument structure of the corresponding simple (i.e. non-prefixed) verb; this relation is realised through the role played by the argument structure of the preverb, provided that we accept the following assumption (Lehmann, 1983, p. 146):
the argument structure of verba composita will be regular just to the degree to which their meaning and constructional behaviour can be explained on the basis of the meaning and constructional behavior of their parts, namely the verbum simplex and the preverb.

In other words, if we limit the analysis to those PVs whose semantics can be predicted from that of the simple verb, such as $a b-e o$ 'come from', then we may analyse the argument structure of the PV in relation with that of the simple verb (eo) and of the preverb ( $a b$-) itself. This also leads us to concentrate on preverbs with a spatial sense, on PVs with compositional meaning and, therefore, on spatial complements occurring as arguments of these PVs.

In the terminology adopted by Lehmann 1983, pp. 146-147 locatum is "something whose position in space or time is described by the construction". Further, relatum is defined as "the item with respect to which the locatum is localized. It is an animate, lifeless or abstract object, a nominal concept which serves as a point of reference for the localization". In general, from a syntactic point of view, the locatum of the preverb is the subject of the PV if the corresponding simple verb is intransitive, and it is its direct object if the simple verb is transitive (Lehmann, 1983, p. 148). In the particular domain of Latin spatial preverbs 'locatum' may be taken as corresponding to what Talmy (2000) calls 'Figure', and 'relatum' to what he calls 'Ground' (cf. section 6.3.3), although these definitions belong into two different theoretical paradigms. In sentence
(13) the locatum coincides with the subject of the verb. ${ }^{16}$ In sentence (12) ortus is the relatum of the preverb ad.

Analogously to sentence (12), double accusatives expressing relata and locata are attested with PVs originating from transitive verbs prefixed with preverbs/prepositions requiring the accusative case: mostly circum- and trans-. ${ }^{17}$ Sentence (14) shows an example of this construction, where the second accusative exercitum (locatum) is the direct object of the verb traducere and the first accusative flumen (relatum) can be seen as governed by the preverb.
flumen Axonam $\quad[\ldots]$ exercitum traducere maturavit
river:ACC.N.SG Axona:ACC.M.SG $[\ldots]$ army:ACC.M.SG lead over:INF.PRS hasten:IND.PPF.3SG
(Caes., B. G., II, 5)
'he hastened to lead his army over the river Aisne'

Moreover, when the preposition corresponding to the preverb requires an indirect case, like $a b$ with the ablative, the relatum can be expressed as the respective bare case, as exemplified in (15).
(15) domina cogis abire mea? (Prop., El., I, IV, 2)
mistress:ABL.F.SG urge:IND.PRS.2SG leave:INF.PRS my:ABL.F.SG
'Do you urge me to leave my mistress?'

All the examples presented so far show morpho-syntactic realisations of relata as NPs with the case required by the preposition corresponding to the preverb; we can say that, in these cases, the preverb acts as an adposition.

Oniga (2005, pp. 218-219) associates these phenomena to the syntactic valency of the corresponding preposition rather than to the semantics of the preverb only, since he observes that these constructions do not occur with preverbs that do not formally correspond to prepositions, like dis- 'un-' (cf. dissuadeo 'dissuade'). Conversely, while discussing the prevalence of these phenomena with the simple ablativus separativus (with $a b-$, de-, ex- and pro-) over the locativus (with in-), Lehmann (1983, p. 153) claims that this may (at least partially) be based on the meaning of the ablative rather than on the preverb itself.

We can claim that the morpho-syntactic realisations of a preverb's relatum can manifest themselves in the following constructions:

1 (CasePrev) as an NP, whose case is that governed by the preposition corresponding to the preverb:
(16) patriosque adit inpiger ortus (Ov., and fatherly:ACC.M.PL go to:IND.PRS.3SG quick:NOM.M.SG sunrise:ACC.M.PL Met., I, 779)

[^53]'he rushed to his father's orient'
castris egressi (Caes., B. G., II, 11, 1)
camp:ABL.N.PL go from:PTCP.NOM.M.PL
'having marched out of the camp'

2 (CaseNonPrev) as an NP, whose case is not the one governed by the preposition corresponding to the preverb:
(18) num tibi aut stultitia accessit aut superat
perhaps you:DAT.SG or folly:NOM.F.SG come:IND.PF.3SG or overcome:IND.PF.3SG
superbia? (Plaut., Amph., 709)
pride:NOM.F.SG
'is it folly that has come to you, or does pride overcome you?'
(19) fontesque avertitur (Verg., Georg., 499)
and spring:ACC.M.PL turn:IND.PRS.3SG.PASS
'he turns away from the springs'

3 (PrepPrev) as a PP, whose preposition corresponds to the preverb:
(20) canem istam a foribus <aliquis>
dog:ACC.F.SG this:ACC.F.SG from door:ABL.F.PL somebody:NOM.SG
abducat face (Plaut., Mostell., 853)
take away:SBJV.PRS.3SG make:IMP.PRS.2SG
'make somebody move the dog away from the door'
(21) e castris Helvetiorum egressi (Caes., B. G., I,
from camp:ABL.N.PL Helventian:GEN.M.PL go from:PTCP.NOM.M.PL 27, 4)
'having marched out of the Helvetians' camp'
4 (PrepNonPrev) as a PP, whose preposition does not correspond to the preverb:
(22) abin e conspectu meo? (Plaut., Amph., 518)
leave:IND.FUT.2SG from sight:ABL.M.SG my:ABL.M.SG
'will you be away from my sight?'
(23) in cuius participationem nullo modo corpus to who:GEN.SG participation:ACC.F.SG any:ABL.M.SG way:ABL.M.SG body:NOM.N.SG adduci potest (Th., S. P. L., II)
draw:INF.PRS.PASS can:IND.PRS.3SG
'to whose participation the body cannot be drawn in any way'

These constructions can be seen as containing different lexicalization patterns of the portion of Path represented by the preverbs.

In section 6.4.2 we will see how these syntactically different behaviours of PVs can be diachronically mirrored in the differences in the status of the preverbs themselves, i. e. when they behave as adpositions or as satellites.

### 6.4.2 Diachronic hypotheses

The distributions of the four constructions exhibited by PVs and described in section 6.4.1 have been diachronically investigated in the literature. In this section we will present some theoretical hypotheses which have been proposed for these constructions. These hypotheses will be statistically tested in chapter 7 .

Sections 6.4.2.1, 6.4.2.2, and 6.4.2.3 present the hypotheses, while at the same time framing a grammaticalization account on the evolution of Latin preverbs. Section 6.4.2.4 contains some final remarks underlining the complexity of the phenomenon at hand and the difficulties of an exclusively diachronic approach to it.

### 6.4.2.1 Construction 1 (CasePrev)

According to some Latin grammars (Ernout and Thomas, 1972; Traina and Bertotti, 1993), spatial complements containing names of cities and small islands are often found without spatial prepositions, just like the nouns domus 'home', rus 'country', locus 'place', pars 'part', litus 'sea-shore', regio 'region', cornu 'horn', 'marry' and adjectives like medius 'middle', imus 'lowest', summus 'highest'. ${ }^{18}$ In addition, special idioms show the same behaviour: ire malam crucem 'go to hell', exsequias ire 'go to a funeral', alicui suppetias ire/(ad-)venire/proficisci 'go/come help someone', animadverto aliquid 'notice something'.

These two facts have been used to hypothesize a connection between the use of construction 1 (CasePrev) with verbs prefixed with spatial preverbs and "a grammaticalization of the accusativus directionis commonly taken by such verbs" (Lehmann, 1983, p. 156). Along this reasoning, construction 1 (CasePrev) has been considered as a result of some specific stylistic choices such as archaisms, Hellenisms and the poetic genre (Ernout and Thomas, 1972; López Moreda, 1998). It has also been diachronically framed into a more general archaic tendency by simple verbs and PVs to encode local complements with bare cases. In other words, construction 1 (CasePrev) has been considered as a relic of an older stage of Latin, when preverbs still kept some semantic and syntactic autonomy from the verb.

According to these premises, we would expect to find this construction significantly more often in archaic writers and authors (particularly in poets) who make use of archaisms:

Hypothesis 6.4.2.1.1. Construction 1 (CasePrev) is significantly more frequent in the archaic age and in works by poets than in the later ages and in prose writers.

### 6.4.2.2 Construction 2 (CaseNonPrev)

The instances of construction 2 (CaseNonPrev) contain certain "transitivisation" cases: in such cases an accusative NP occurs as the spatial argument of a PV, while the simple verb is intransitive and the preposition corresponding to the preverb does not govern the accusative case (Coleman, 1991, p. 335). For instance, evado 'come out' is prefixed with ex- 'from' which requires the ablative case, but evado can occur with the

[^54]accusative. The following sentence shows an example with the verb averto 'turn away' $<a b$-verto followed by the accusative fontes 'springs'.
fontesque avertitur (Vergil, Geor. 3, 499)
and spring:ACC.F.PL turn:IND.PRS.3SG.PASS
'he turns away from the springs'.

In section 6.4.2.1 we presented some diachronic hypotheses which assigned construction 1 (CasePrev) to archaic and poetic language. The use in the poetic language applies to construction 2 (CaseNonPrev) as well, if we limit the analysis to these "transitivisation" cases, where the accusative is the bare case used for expressing the preverb's relatum (Rosén, 1999).

In addition, a special remark should devoted to the dative case. According to many scholars (cf. e.g. Sweetser 1988, p. 395, Coleman 1991, p. 335, Bortolussi 2005), in several languages - among which Latin - a spatial metaphor drives the choice between dative and ablative in local complements, according to which dative is employed for Goal expressions and ablative for Source.

Regarding the evolution of the contructions involving the dative, Rohlfs (1970, pp. 6-7) notices that in Latin - and later in Romance languages - ad 'to' + accusative can be used to replace a dative case, and Vincent (1999, pp. 1115-1116; 1130-1131) distinguishes this case-marking use of $a d$ from a use of prepositional $a d$. This fact should be taken into account while considering construction 2 (CaseNonPrev), since a motion PV prefixed with $a d$ - can appear with the bare dative case to express direction. Furthermore, Rosén (1999, p. 146) argues that the so-called "dative of direction", used instead of a PP to express the direction within a spatial complement, is preferred by Augustan poets. In addition, the use of the dative as opposed to a PP is considered to be more frequent when the verb has a figurative sense and the complement position is lexically filled by an animate entity, especially when the subject is inanimate (Longrée, 2005, p. 305). Longrée (2005) also notices that some idiomatic expressions - like timor/terror incessit+dative 'fear/terror struck' +NP - might have favoured analogical constructions - such as terror incidit+dative 'terror fell upon'+NP. Finally, Pinkster (2010) analyses some PVs occurring with dative complements and finds that this construction "can be explained by referring to one or more (classes of) verbs with a similar meaning, or especially in creative texts by assuming a more complicated bridge, which is non the less semantic". It is thus realistic to assume semantic factors that may explain the use of construction 2 (CaseNonPrev) with dative.

In conclusion, the examples of construction 2 (CaseNonPrev) are difficult to categorize from a diachronic point of view, given that different hypotheses have been proposed depending on the specific case involved (accusative or dative). Since a specific analysis on these constructions falls outside the scope of our work, the investigations in chapter 7 will aim at testing a general preference for this construction in poets.

Incorporated and lexicalized relational preverbs While analyzing preverbs in Homeric Greek, Imbert (2008) distinguishes between incorporated relational preverbs and lexicalized relational preverbs. Both these classes syntactically occur with the accusative case even if the corresponding prepositions do not govern this case. According to these definitions, incorporated relational preverbs are not semantically lexicalized within the PV: cf. ekdiabaínō 'go through from' followed by an accusative NP, although $e k$ - 'from' governs the genitive case.

Incorporated relational preverbs share some properties with applicative markers studied in typology (Viti, 2008; Imbert, 2008). Applicatives are defined by Peterson
(2007, p. 1) as "a means some languages have for structuring clauses which allow for the coding of a thematically peripheral argument or adjunct as a core-object argument". Here we report one of his examples from Ainu, a language spoken in the Japanese island of Hokkaido, where the verbal prefix $e$ - assigns the locative function to the bare noun cise.
(25) poro cise e-horari
big house live:APP
'He lives in a big house'.
On the other hand, Imbert (2008) argues that lexicalized relational preverbs have undergone some degree of lexicalization and loss of semantic compositionality. An example is the Ancient Greek verb apanaínomai 'deny', prefixed with apó- 'off' (relational preverb, requiring the genitive) and occurring with an accusative object. Furthermore, according to Imbert's categorisation, Goal preverbs almost never lexicalize in Homeric Greek, while Source preverbs tend to become lexicalized relational preverbs more easily. If we extend Imbert (2008)'s definition of incorporated and lexicalized relational preverbs to PVs occurring with bare cases different from the accusative, construction 2 (CaseNonPrev) corresponds to these groups of preverbs.

### 6.4.2.3 Constructions 3 (PrepPrev) and 4 (PrepNonPrev)

Construction 1 (CasePrev) involving the bare case has been generally recognised as being less innovative than constructions 3 (PrepPrev) and 4 (PrepNonPrev), which involve the presence of a preposition (Coleman 1991, p. 332, Rosén 1999, p. 37).

This appearance of innovative constructions co-existing with older ones could be interpreted in the light of the diachronic development of preverbs from an adpositionallike behaviour to a progressive lexicalization, which was discussed in section 6.3.2.2. Indeed, construction 1 (CasePrev) may be viewed as an instance of the more "original" situation, where preverbs acted as adpositions, thus syntactically governing the case of the NP expressing their relata (section 6.3.3.2). Conversely, the presence of the preposition in constructions 3 (PrepPrev) and 4 (PrepNonPrev) could be connected with an innovative situation, where the gradual lexicalization of preverbs within the verbs and the increasing degree of semantic non-compositionality of the PVs might have favoured the use of PPs to express spatial complements. The decreasing proportion of construction 1 (CasePrev) and the increasing proportion of constructions 3 (PrepPrev) and 4 (PrepNonPrev) with time would thus be in accordance with this evolution. This is, of course, one of the possible perspectives from which we can examine the distribution of these constructions. It is also possible that stylistic and pragmatic factors drove the author's syntactic choices. However, if we only focus on a diachronic perspective, we can formulate the following hypothesis:

Hypothesis 6.4.2.3.1. The frequency of construction 1 (CasePrev) with respect to constructions 3 (PrepPrev) and 4 (PrepNonPrev) tends to decrease with time.

If we consider the literary genre rather than chronology only, we can formulate some additional hypotheses. López Moreda (1998, p. 964) claims that construction 3 (PrepPrev) is more frequent in the dramatic genre and in the familiar and colloquial register. Rosén (1999, pp. 148-149) identifies stylistic criteria affecting the choice between the different constructions in the later Republican and early Imperial age, associating the use of PPs with the familiar register and the use of bare case to a more elaborated style. Summing up, we will hypothesize a general preference for construction

3 (PrepPrev) in dramatic writers and authors whose style contains some features from the spoken language.

Preverbs as satellites In constructions 3 (PrepPrev) and 4 (PrepNonPrev) we may have instances of redundancy (Sinha and Kuteva, 1995), because the Path component may be at the same time expressed by the preverb and by the preposition.

```
(26) e castris [..] egressi (Caes., B. G., I, 27, 4)
    from camp:ABL.N.PL go from:PTCP.NOM.M.PL
    Path Ground Translational Motion + Path
    'having marched out of the Helvetians camp'
(27)
    abin ee conspectu [...]? (Plaut., Amph., 518)
    go from:IND.FUT.2SG from sight:ABL.M.SG
    Translational Motion+Path Path Ground
    'will you be away from my sight?'
```

As a matter of fact, it is difficult to detect a weakened transparency of the preverb in the verb's semantics and decide whether the Path component is actually expressed in the preverb or not: in sentence (27), for example, the verb $a b$-eo could have either kept the compositional meaning 'go from', or lost it, simply meaning 'go, leave'. Although the verb's semantics is subject to interpretation, we can nevertheless hypothesize that the presence of the prepositions $e$ 'from' in (26) and (27) is related to the fact that the preverb played a role which is different from the adpositional role.

Imbert (2008) calls these preverbs satellite preverbs and claims that they represent the end of a grammaticalization process that started from relational preverbs. Moreover, Imbert refers to the semantic-conceptual criterion of the Goal/Source asymmetry. The Goal/Source asymmetry has been investigated in several subfields of linguistics, including cognitive linguistics and typology. Among the typological studies on syntax, Ikegami (1987) analyses the asymmetry in the way languages express Source and Goal, showing psychological reasons for what he calls the 'goal-over-source principle'. According to this principle, the expressions of Goal are predominant in languages if compared with expressions of Source; evidence for this can be found in the possibility of replacing the Source marker with the Goal marker, and in the markedness of Source expressions. Among the psycholinguistic studies, Lakusta and Landau (2005) show that there is a non-linguistic and linguistic bias towards Goal in encoding paths; this bias applies to all event types and is evident in both children and adults.

According to Imbert, the Goal/Source asymmetry partially drove the process of grammaticalization for Ancient Greek and Old English preverbs: Goal preverbs tended to hold an adpositional role (relational preverbs), while Source preverbs tended to lexicalize (satellite preverbs). In our analyses we will investigate to what extent the Goal/Source nature of preverbs affects the type of constructions occurring with their PVs.

### 6.4.2.4 A few more remarks

The general trend which took place throughout the history of Latin from the IndoEuropean origins to Romance languages and which led to lexicalization of certain Latin preverbs can only be detected over a large time span. Consequently, the effects of the trend are not necessarily "measurable" at each point of this process. In other words, the developments described in the previous sections do not seem to explain the whole story about the interplay of synchronic and diachronic factors governing the expression of locata and relata of Latin preverbs. As a matter of fact, it is possible to find both the presence and the absence of prepositional constructions in the same author, already
in the archaic age (Oniga, 2005); this is sometimes justified by certain stylistic choices (Ernout and Thomas, 1972, p. 33) and by the principle of variatio (Longrée, 2005, p. 302 ), or even by pragmatic factors. In addition, the lexicalization of preverbs is not only a chronological feature relative to a particular historical age, but is attested since early Latin.

The phenomenon is in fact more complex than a general trend would suggest, since the preference towards a particular construction is prone to influence of additional criteria, involving for example the animacy and lexical properties of the locatum and relatum, as well as the particular (concrete or figurative) sense of the verb (Longrée, 2005; Viti, 2008).

In the corpus-based studies reported in chapter 7 we focus on the four constructions occurring with Latin PVs, aiming at finding the conditions under which one construction is preferred over the others. Our general assumption is that there is a connection between a more advanced degree of lexicalization of the preverbs within the PVs and a preference for prepositional constructions such as construction 3 (PrepPrev) and 4 (PrepNoPrev). In particular, in section 7.4 we will test whether this can be associated with a general diachronic trend and how it can be detected in corpus data, while in section 7.5 we will show how variables other than the chronological one should be taken into account in the analysis.

## Chapter 7

## A corpus-based study on Latin spatial preverbs

### 7.1 Introduction

This chapter presents a case-study on Latin preverbs: in particular, we investigate the argument structure of Latin verbs which are prefixed with a spatial preverb.

As introduced in chapter 6, preverbation traditionally pertains to word formation and is therefore a morphological phenomenon; however, preverbed verbs (PVs) give us the chance to explore interesting aspects of the syntax-semantics and syntax-lexicon interfaces. We will show how relevant interactions between the morpho-syntactic realisations of the arguments of PVs and their lexical-semantic properties can diachronically manifest themselves and be detected in corpus data.

The new contribution of our work is mainly methodological and addresses a common need for empirical investigation on this topic. The source data for our corpus-based analysis is the subcategorization (SC) lexicon described in chapter 4 and two plays by Plautus: Amphitruo and Mostellaria.

The rest of the chapter is organized as follows. Section 7.2 outlines the design of our research, while section 7.3 illustrates the corpora and the dataset used for our analyses. Finally, sections $7.4,7.5$ and 7.6 present our corpus-based quantitative analyses and section 7.7 provides the conclusions.

### 7.2 Research design

This section briefly states the goal of our research, and illustrates the empirical approach we followed, for which a more general introduction is provided in section 1.3.

Goal and approach The linguistic goal of this research is to carry out a corpus-based quantitative analysis on the argument structure of Latin PVs. In addition, we aim at proposing a more general empirical approach which can be adapted to other linguistic phenomena in Latin.

The approach we chose to achieve our linguistic goal is quantitative and contrasts with the qualitative approach followed in all previous studies on Latin PVs. This means
that the corpora are not used to provide a set of examples displaying the evidence for theoretical statements, as in qualitative approaches; instead, they constitute the data to be systematically analysed through specific statistical techniques.

Our procedure starts from collecting the data with relevant linguistic information (variables) in the form of corpus instances of PVs (observations); part of this information was originally present in the annotated corpora (morphological and syntactic features), and the rest was added manually (verb sense, verb class and animacy annotation).

On the one hand we will follow the deductive reasoning typical of confirmatory data analysis and inferential statistics, on the other hand, we will employ an exploratory approach.

Confirmatory data analysis As summarised in various works on this topic (cf., for example, Rasinger 2008, pp. 10-18), the deductive reasoning behind confirmatory data analysis (also called hypothesis testing) starts from a theory in the form of a hypothesis (model): for instance the freer word order in archaic Latin as opposed to a more fixed order in classical Latin. The next step involves developing an appropriate methodology that produces quantitative data. Typically, the data are built so that there is one variable under investigation (e.g. some index of word order) and the procedure functions as a controlled experiment, where all the other factors are kept constant in order to detect a possible change in that single variable. Then, the aim of the research is to prove or disprove the hypothesis by drawing inferences from the sample (corpus data) to the population (a definition of the sublanguage we want to model) in an inferential perspective.

Adopting this approach requires a clearly stated hypothesis to test. As we saw in section 6.4.2, in our case some general hypotheses can be formulated and will be tested here. However, the complexity of the phenomenon is so great that there are several variables contributing to it and it is not always possible to formulate theoretical hypotheses in advance. For this reason, we will also employ a different approach.

Exploratory data analysis This term was coined by Tukey (1977) and indicates an approach to data analysis that aims at uncovering the underlying structure of the data.

Unlike confirmatory data analysis, this view requires the researcher to explore the data, open to gain new insight into it. The model then "emerges" from the data rather than being confirmed or not by the data. This view is very suitable when a high degree of variability is present in the data and the researcher does not have a pre-defined theory in mind.

After a first deductive investigation, we will adopt an exploratory approach in order to fully account for the variability in the data. This way, we will not "impose" a theoretical view on the data, but we will let the analysis produce the correct model fitting the data at hand, thus letting relevant interactions between the variables emerge from the analysed data.

For most statistical analyses, we will make use of the interactive programming environment R (R Development Core Team, 2008). In addition, Multiple Correspondence Analysis will be performed using the $c a$ package in R and the MATLAB toolbox Analytica (Apollon, 2003).

### 7.3 The corpora and the dataset

In this section we present the corpora (section 7.3.1) and the dataset (section 7.3.2) we used for the analysis.

### 7.3.1 Corpora

In the present section we describe the composition of the three corpora which are the base for our study.

We decided to collect the corpora for our research so that they were as extensive as possible, while at the same time requiring as little manual pre-processing as possible. For this, the SC lexicon described in chapter 4 proved to be the optimal choice: it is the only corpus-driven lexicon for Latin containing information on verbal arguments and their morphological, syntactic and lexical realisations. Being corpus-based, it contains the frequency information necessary for our analysis. In order to broaden the chronological and genre range of the authors considered, we also manually analysed two plays by Plautus. Therefore, the basis for our empirical research consisted of three datasets:

- Plautus: Amphitruo and Mostellaria. ${ }^{1}$
- Latin Dependency treebank (LDT; cf. section 3.2.3.1, page 40):
- Caesar: selections from book II of De bello gallico;
- Cicero: In Catilinam, 1.1-2.11;
- Sallust: Catilina;
- Ovid: Metamorphoses, book I;
- Propertius: Elegies, book I;
- Vergil: Aeneid, selections from book VI;
- Petronius: Satyricon 26-78 (Cena Trimalchionis);
- Jerome: Vulgate (Apocalypse).
- Index Thomisticus treebank (IT-TB; cf. section 3.2.3.1, page 40):
- Thomas Aquinas: selections from Scriptum super Sententiis Magistri Petri Lombardi;
- Thomas Aquinas: selections from Summa contra Gentiles.

Good corpora Section 1.3 discussed the reasons why we can consider our corpora as good corpora for the study at hand. Here we will summarise them.

First, a pragmatic reason: this is the only (and thus the best) explicit corpus evidence we have for the argument structure of Latin PVs.

Second, the corpora were collected without a particular study on PVs in mind: in this sense, they can be defined as general purpose corpora. This enables us to exclude the presence of biases and ensures their being a representative sample of our population. Therefore, they can be considered a random sample of the Latin sublanguage we operationally defined in section 1.3.0.1.

Third, the chronological span covered by these corpora allows for a diachronic approach in the study we plan to conduct.

[^55]
### 7.3.2 Dataset

This section presents the actual dataset on which our analysis was carried out. The criteria of preverb selection and verb selection are given in sections 7.3.2.1 and 7.3.2.2. According to these criteria, an appropriate portion of all instances of PVs in the corpora listed above was extracted (section 7.3.2.3): this subset is the dataset for our study.

### 7.3.2.1 Selecting the preverbs

We focussed on a list of preverbs which also have a spatial meaning: ${ }^{2}$

- ab- 'away from';
- ad- 'to, toward; at, by, on';
- ante- 'before, in front of, forwards';
- circum- 'around';
- de- 'from, down from, off';
- ex- 'out of';
- in- 'in within, on, into, to';
- inter- 'between, among, through';
- intro- 'within, inside of';
- ob- 'towards, against; before';
- per- 'through';
- prae- 'before', 'in front of';
- pro- 'before', 'in front of';
- sub- 'under';
- trans- 'across, through, over'.

With respect to the semantics of the preverbs, a reminder of the polysemy of most Latin preverbs is in order here. About this issue, García-Hernández (2005, pp. 230-235) notices the semantic evolution of spatial preverbs from a concrete to a more abstract meaning. Consider, for example, the allative preverbs $a d$ - and $i n$-, which originally indicated a direction in space, and then gained an abstract meaning: cf. the two verbs adeo 'go to, approach' and ineo 'go into, enter', which also acquired the ingressive metaphorical senses of 'undertake something' and 'start something', respectively. Moreover, ablative preverbs that indicate the direction from a place - like $a b-$, $e x$ - and $d e--$ had an "aspectual evolution towards resultative": cf. sumo 'take' vs. absumo 'take away, consume' and bibo 'drink' vs. ebibo 'drink up'. Finally, the preverbs per- (longitudinal direction) and trans- (transversal direction) had "an evolution toward progressive aspect": cf. the verbs perago and transigo which mean 'accomplish' vs. ago 'put, lead'.

[^56]In our analysis we only focussed on the instances of these preverbs where they had a spatial meaning. Additionally, we left out those preverbs that have an aspectual or intensive meaning rather than a spatial one, such as per- in per-venio 'arrive at', which is telic unlike venio, and does not compositionally mean 'go through'.

The lexicon The starting point of our analysis is the set of all verb lemmas in our SC lexicon: this was made possible by the information on PoS and lemma contained in the treebanks. Then, we selected those PVs occurring in the corpora and prefixed with the previously listed preverbs. ${ }^{3}$

Plautus From the two plays by Plautus we manually extracted all occurrences of PVs whose preverbs belonged to the previous list of spatial preverbs.

This manual procedure was chosen because no treebank containing Plautus' works is available yet $;^{4}$ we considered that a manual annotation of the arguments of some specific verbs (PVs) in two plays was quicker than building a treebank for this purpose. Further, no parser for Latin has been trained on Plautus nor on writers from the early republican age. ${ }^{5}$ Given the peculiarities of Plautus' style and the age he belongs to, we hypothesized that the accuracy of a parser on these texts would have been low.

### 7.3.2.2 Selecting the verbs

We selected those PVs occurring in the corpora that met the following additional requirements:

- Semantics of the PV:
- verbs of motion: e.g. adeo 'go to', 'reach';
- verbs of rest: e.g. adsum 'be present';
- verbs of transport: e. g. adduco 'lead to'. ${ }^{6}$
- Semantics of the PV in relation with the semantics of the preverb and the semantics of the simple verb: the meaning of the PV is compositional: e.g. insum 'be in'. The semantics of the PVs was analysed by inspecting the verbal occurrence in context and the verb senses listed in the entries from the Perseus Latin Dictionary Online (Lewis and Short, 1879). Thus, we categorized the verbs on the basis of the following criteria:
- the PV has a concrete locative sense: abducere aliquem in agros 'lead somebody into the fields', abire in terras 'go towards the countries';
- the PV has a figurative (metaphorically locative) sense: abire in villos 'be transformed in a tuft of hair', absolvere de coniuratio 'loosen from the conspiracy';

[^57]- the PV has an abstract non-locative sense: promittere alicui aliquid 'promise something to somebody'.

The second category was distinguished from the first one by the label 'figurative', while the last case was excluded from the analysis.

- Frequency: the PV has a type frequency greater than 2. This frequency threshold was set in order to exclude idiosyncratic usages that may not reflect a reliable evidence for our analysis, or simply annotation errors.


### 7.3.2.3 Extracting the verbal arguments

The goal of our study concerns the morpho-syntactic realisations of the arguments of PVs. More specifically, since the preverbs we selected have a spatial meaning, we are interested in the realisation of spatial arguments of PVs, in terms of relatum and locatum of the preverb, and other locative complements, such as Goal or Source expressions (see section 6.3.3). As Brachet (2000, pp. 10-11) points out, the phenomenon of preverbation is said to turn adjuncts into arguments. Along the same lines, Viti (2008, pp. 381382) claims that "Complement-taking preverbs are much rarer than prepositions [...] This suggests that case-assigning preverbation is a derived strategy with respect to preposition, and is used to promote an adjunct to an argument". We collected both the instances of case-assigning preverbs and PPs occurring with PVs and expressing the preverbs' locata; such complements are annotated as arguments in the treebanks and therefore appear in the SC lexicon. Consequently, we considered them as part of the PVs' argument structure.

Since a manual inspection of the single selected occurrences was necessary in order to annotate the semantic features of the verb and the arguments, we also corrected some errors present in the lemmatization, as well as in the morphological and syntactic tagging.

Table 7.1 displays the composition of the corpora in terms of the number of verb tokens per author and (in bold) the number of PVs observed in each author: the last figures correspond to the actual size of our dataset. ${ }^{7}$ The number of verbal tokens for Plautus is not displayed, since the extraction of PVs was done by manual inspection and we do not have a list of all verbs occurring in the two plays by Plautus.

### 7.4 First investigation

In this section we report on the first corpus-based analysis we carried out in order to test the hypotheses formulated in section 6.4.2. In this first exploration we focussed on the realisation of preverbs' relata, that is on the realisation of the portion of motion events represented by the preverbs. For example, when a preverb expresses the Source of a motion event (like $a b-e o$ 'go from') we collected the Source complements occurring with that verb, for example abeo domo lit. 'go from home', 'leave home', abeo a te 'go away from you', or abeo e conspectu 'go away from the sight'.

The present section describes the dataset (section 7.4.1) and presents the statistical analysis (sections 7.4.2 and 7.4.3) for this first study, and finally contains concluding remarks (section 7.4.4).

[^58]| corpus | author | verb tokens | PV tokens |
| ---: | :--- | :---: | :---: |
| Plautus | Plautus |  | $\mathbf{1 1 4}$ |
|  | Cicero | 1283 | $\mathbf{3 0}$ |
|  | Caesar | 295 | $\mathbf{1 3}$ |
|  | Sallust | 2295 | $\mathbf{5 2}$ |
| LDT | Vergil | 511 | $\mathbf{2 7}$ |
|  | Propertius | 1039 | $\mathbf{2 1}$ |
|  | Ovid | 1092 | $\mathbf{2 3}$ |
|  | Petronius | 2901 | $\mathbf{6 9}$ |
|  | Jerome | 1770 | $\mathbf{4 8}$ |
| IT | Thomas Aquinas | 9072 | $\mathbf{1 5 0}$ |
|  | TOTAL |  | $\mathbf{5 4 5}$ |

Table 7.1: Number of word tokens, verb tokens and PV tokens (in bold) by author in the data. The number of verb tokens for Plautus is not recorded because the original data are not tagged by part of speech.

### 7.4.1 Dataset

In this study we looked for a diachronic trend (if it existed) in the opposition between some of the constructions illustrated on page 132. In particular we focussed on the alternation between construction 1 (CasePrev) on the one hand, and constructions 3 (PrepPrev) and 4 (PrepNoPrev) on the other, i.e. between the presence of the preposition and the presence of the NP (bare case) governed by the preverb in the realisation of the relatum of the preverb. ${ }^{8}$

Consequently, we limited ourselves to the instances where the relatum of the preverb fell in one of the following cases:

- Construction 1 (CasePrev): an NP whose case coincides with that governed by the preverb: exercitum Ligerim traducit 'he leads his army across the Loire' (Caes., B. G. $7,11,9)$;
- Construction 3 (PrepPrev): a PP whose preposition corresponds to the preverb: inter me atque te murus intersit 'there is a wall between me and you' (Cic., Cat. I, 5);
- Construction 4 (PrepNoPrev): a PP whose preposition does not correspond to the preverb: abin e conspectu meo? 'will you be away from my sight?' (Plaut., Amph., 518).

In order to represent this alternation we set a binary variable called 'construction'; construction 1 (CasePrev) listed above is identifiable in a univocal way by the value 'CasePrev' of this variable, while constructions 3 (PrepPrev) and 4 (PrepNoPrev) correspond to the value 'Preposition'.

Whit these selection criteria we extracted 237 verbal tokens, which constitute the first dataset.

[^59]| author | CasePrev | Preposition |
| :--- | :---: | :---: |
| Plautus | 12 | 41 |
| Cicero | 1 | 7 |
| Caesar | 1 | 5 |
| Sallust | 7 | 13 |
| Vergil | 7 | 0 |
| Propertius | 5 | 2 |
| Ovid | 5 | 1 |
| Petronius | 9 | 18 |
| Jerome | 0 | 29 |
| Thomas | 0 | 74 |

Table 7.2: Frequency counts of 'CasePrev' and 'Preposition' by author.

### 7.4.2 Statistical tests

We first decided to look for a general trend in our dataset for what concerns the distribution of the binary variable 'construction'.

The most fine-grained diachronic categorization in our data is by author; hence, 'author' was chosen as the second variable under study. Table 7.2 contains the distribution of 'construction' in the different authors attested in the dataset, while figure 7.1 is a barplot displaying the proportion of the two values of 'construction' ('CasePrev' and 'Preposition') over all PV tokens by author. Figure 7.1 shows that the proportion of construction 'CasePrev' is particularly high in the poets (Ovid, Propertius and Vergil) than both the earlier and the later authors.

Figure 7.2 plots the absolute frequencies of the two values for 'construction' in each author, while figure 7.3 plots the relative frequencies. From these figures we can observe a general trend for the presence of prepositions with respect to their absence. The dotted line ('CasePrev') remains under the solid line ('Preposition') in all authors except for the poets (Vergil, Propertius, Ovid). The proportion for Plautus, Cicero, Caesar are similar to the proportions for Jerome and Thomas.

A simple comparison of relative frequencies may lead to misinterpretations, especially when dealing with low figures. For this reason, we will investigate these data from a statistical point of view, in order to detect a significant difference appearing in the distribution of these frequencies. We will then give a quantitative account for the interaction between the two variables 'author' and 'construction' by referring to hypothesis testing introduced on page 140: we will test Hypothesis 6.4.2.1.1 on page 134 and Hypothesis 6.4.2.3.1 on page 136 against the Null Hypothesis 7.4.2.0.1.

Null Hypothesis 7.4.2.0.1. In the first dataset the variables 'author' and 'construction' are independent from each other.

We used the Pearson $\chi^{2}$ test for independence illustrated in Appendix C.2, which allows us to decide whether we are able to reject this null hypothesis, given the data at hand. A $\chi^{2}$ test works by comparing the actual frequencies observed in the data with the expected frequencies we would get from a random distribution. This test gives the level of significance as output, which states whether the result is significant or not, i. e. if the two variables are significantly independent or not. In addition, it is important to calculate the effect size, which accounts for the size of the detected dependence.


Figure 7.1: Barplot of the distribution of 'CasePrev' (light grey bars) and 'Preposition' (dark grey bars) by author; the percentage over all verbal tokens by author is displayed.


Figure 7.2: The variables 'author' and 'construction' plotted against each other in the first dataset (absolute frequencies).


Figure 7.3: The variables 'author' and 'construction' plotted against each other in the first dataset. The proportions of the two values of 'construction' in each author are displayed.


Figure 7.4: Association plot for the $\chi^{2}$ statistics on the two values for the variable 'construction' by author in the first dataset.

The results from the independence test on the data from table 7.2 reveal that the variable 'author' and the variable 'construction' are not independent at the 0.01 level of significance $\left(p<0.01, \chi_{(d f=9)}^{2}=87.28\right)$, so we can reject the null hypothesis 7.4.2.0.1. The strength of this association is large $(V=0.61)$ according to the classification in Cohen (1988, p. 224)

We now know that somewhere in our data there is a significant difference in the distribution of the two constructions 'CasePrev' and 'Preposition'. If we want to see where this difference is concentrated, we can inspect an association plot. Figure 7.4 is a Cohen-Friendly association plot and shows the contribution made by each authorconstruction pair - i.e. each cell of table 7.2 - to the total $\chi^{2}$ statistic. In this plot the area of the each rectangle is proportional to the difference between observed and expected frequencies, where the dotted lines refer to expected frequencies. The height of the rectangles is related to the significance and the width to the amount of data that support the result.

From figure 7.4 we notice that in the data for Thomas and Jerome the construction 'CasePrev' occurs significantly less frequently than expected, whereas construction 'Preposition' occurs more frequently than expected; on the other hand, Sallust, Petronius and the poets - Ovid, Propertius and Vergil - display the opposite behaviour. Note that a simple inspection of the barplot in Figure 7.1 for Sallust and Petronius would have erroneously implied the contrary. In the case of Plautus' plays, we seem to have a rather large data evidence, but the deviation of the frequencies of the two constructions from the expected frequencies is small. This may imply that the variability in these data can be captured by a different and more complex combination of variables: see

| author | CasePrev | Preposition |
| :--- | :---: | :---: |
| Plautus | 10.51 | 42.49 |
| Cicero | 1.59 | 6.41 |
| Caesar | 1.19 | 4.81 |
| Sallust | 3.97 | 16.03 |
| Vergil | 1.39 | 5.61 |
| Propertius | 1.39 | 5.61 |
| Ovid | 1.19 | 4.81 |
| Petronius | 5.35 | 21.65 |
| Jerome | 5.75 | 23.25 |
| Thomas | 14.68 | 59.32 |

Table 7.3: Expected frequencies corresponding to the frequencies in table 7.2.
the study reported in section 7.4.3. Due to the low frequencies for Caesar and Cicero, these authors contribute very little to the overall association between the two variables, which makes it difficult to generalize from the data for these authors.

As a matter of fact, the $\chi^{2}$ test gives unreliable results when applied to a table whose expected frequencies are lower than 5 . Table 7.2 produces 8 out of 20 expected cell frequencies below this recommended threshold, as shown in table 7.3. For this reason, we merged together the frequencies for the poets (Ovid, Propertius and Vergil) on the one hand, and the frequencies for the late republic authors (Caesar, Cicero and Sallust) on the other hand. This way, only one cell (corresponding to the construction 'CasePrev' for the poets) of the 20 expected frequencies is lower than 5. A $\chi^{2}$ test performed on these data still resulted in a significantly large association between 'author' and 'construction' $\left(p<0.01, \chi_{(5)}=83.20, V=0.59\right)$, although this procedure causes a loss of information during the interpretation of the results because it obviously does not allow us to make distinctions between the single authors merged.

The association plot for these data is represented in figure 7.5. These preliminary results confirm the claims given in section 6.4 .2 on the preference of poets and the earliest author in our corpus - Plautus - for this construction (Hypothesis 6.4.2.1.1 on page 134). If we look for a progression in time, we realise that there is not a one-way pattern emerging from our data. In fact, figure 7.5 shows a decrease in the distribution of 'CasePrev' - and a symmetrical increase for 'Preposition' - from the poets to Petronius and, later, Jerome and Thomas. However, the first two chronological groups (Plautus and the late republic authors) do not fit this trend. One of the reasons for this may be a "structural" feature of our data, namely the fact that the chronological variable is in our case not distinguishable from the genre variable, as attested by table 1.2 on page 8 . This places the late republic authors among the prose writers and the Augustan writers among the poets; we would need data from late republic poets and Augustan prose writers to see if the detected trend is a chronological one, or if it is instead related to genre. In a similar vein, the data from the late imperial age (Jerome) and middle ages (Thomas) overlap with the data for two kinds of religious texts.

Since there is no way to solve this variable overlap, we will try to achieve the most informative results from the data at hand. If we group the authors in two main eras 'early' (Plautus, Cicero, Caesar, Sallust, Vergil, Propertius, Ovid and Petronius) and 'late' (Jerome and Thomas) - we obtain a statistically significant, medium-to-large difference in the distribution of the two constructions 'CasePrev' and 'Preposition'


Figure 7.5: Association plot for the $\chi^{2}$ statistics on the two values for the variable 'construction' by author in the first dataset. We merged the data for Caesar, Cicero and Sallust under the label 'late republic', and the data for Ovid, Propertius and Vergil under the label 'poets'.
$\left(p<0.01, \chi_{(1)}^{2}=42.88, V=0.44\right)$. This results in a clear trend, which constrasts the early age and CasePrev on the one hand, and the late age and Preposition on the other, thus corroborating Hypothesis 6.4.2.3.1.

### 7.4.3 A multivariate approach

The first investigation we reported on in section 7.4 could only account for the interaction between a diachronic variable - 'author' - and our "target" variable - i. e. the syntactic realisation of the preverb's relatum. In this context, a $\chi^{2}$ test was able to detect some differences in the use of one construction (CasePrev) over another (Preposition) by certain authors. In some respects, this result can be considered a diachronic trend in the data.

However, the aim of this study is not only to verify whether the two alternative constructions are used in a significantly different way among the authors, but also to detect the relevant variables that influence this difference, such as verb semantics, animacy of the fillers of the arguments, etc.

We conducted a $\chi^{2}$ test for the association between our "target" variable 'construction' and the variable 'preverb'. The result is a significant medium-to-large association $\left(p<0.01, \chi_{(8)}=50.38, V=0.46\right)$, slightly lower than the effect size found for 'construction' vs. 'author' ( $V=0.61$, see page 149), but still considerable. This shows us that both the variables 'author' and 'preverb' are relevant for our study, since they both affect the realisation of the preverb's relatum as encoded by the variable 'construction'. Consequently, if we want to explore the correlation between our target variable and more than one variable - for example the type of preverb involved, as well as the author - we need more advanced techniques. In addition, we aim at using techniques that allow us to pinpoint or exclude interactions and effects ascribable to each variable.

In this section we will report on a second study conducted on the first dataset by using regression modeling, illustrated in section C.3. Linear regression models aim at quantifying the relationship between a specific variable - called response- and a set of independent variables - called predictors. Here we will make use of a particular type of linear regression models: the generalized mixed effect models (cf. Appendix C.3.4). The peculiarity of these models consists in the random effects they contain. The random effects allow us to (at least partially) compensate for biases in the data that may make the sample not representative enough of the population.

We saw that our dataset has an uneven composition as far as chronological ages and genres are concerned (see the discussion in section 1.3); this reflects the selection of authors and cannot be eliminated if not at cost of considerably decreasing its size. However, in the mixed effect models we considered 'author' as one of the random effects: this way, we assume that the dataset consists of a collection of random samples from an unknown, potentially infinite population of authors (see section C.3.5 for the details of the model), instead of a single sample from one population of authors.

### 7.4.3.1 New variables

We aimed at quantifying the relationship between 'construction' - taken as the so-called 'response' - and five variables, called 'predictors':

- the type of preverb the PV is prefixed with (variable 'preverb');
- a diachronic variable for the literary age: 'era', defined in table 1.2 on page 8 and taking the values 'early_republican', 'late_republican', 'Augustan', 'imperial',
'late_imperial' and 'medieval';
- a classification of the preverbs based on the case required by the corresponding preposition: variable 'case', values 'accusative' as for $a d-$ - 'to' and 'ablative' as for $a b$ - 'from';
- 'animacy_relatum': 'animate' if the preverb's relatum is animate, 'non_animate' otherwise;
- 'figurative', encoding whether the PV is used with a figurative sense or not.

The annotation of both 'animacy_relatum' and 'figurative' was performed manually, since it was not present in the treebanks. We annotated each instance of a PV from the dataset in its context, thus identifying the animacy of the fillers (including the cases when the filler is a pronoun) and figurative verb usages. This type of annotation can be connected with the verbs' selectional preferences dealt with in chapter 5, where lexicalsemantic properties of the verbal arguments are projected on the arguments' fillers via an external ontology such as WordNet.

The random effects are 'author' and 'class', i. e. the verb class of the PV, which is either 'motion' (e.g. abeo 'leave'), or 'rest' (e.g. adeo 'be present'), or 'transport' (e.g. abduco 'lead from').

### 7.4.3.2 The linear model

The model is hence:

$$
\text { construction } \sim \text { preverb }+ \text { era }+ \text { case }+ \text { animacy_relatum }+ \text { figurative }+(\text { class } \mid \text { author })
$$

where we indicated the response, followed by the predictors and the random effects.
After trying several models involving different combinations of the variables, we found that this model was the best one for predicting the values of 'construction' in our dataset. This means that the value of 'construction' can be best predicted as a linear function of the type of preverb, the literary age, the case governed by the preposition corresponding to the preverb, the animacy of the relatum's lexical filler, the figurative use of the verb and by grouping together the data on the basis of the author and the broad verb class. Appendix C.3.5 reports the output and the diagnostic plot. The obtained model estimates the probabilities that the construction is 'preposition' while switching from the reference level (early republican era, preverb $a b$-, ablative case, non animate relatum, non figurative verb usage) to the value of each predictor, as shown in table C. 4 on page 242.

On the whole, the preverbs have a small (negative with respect to the reference level) effect on the type of construction, the exceptions being $a b-(81 \%$, reference level), de- ( $42 \%$ ) and ex- 'from' ( $58 \%$ ): the probability of 'construction' having the value 'preposition' is $10 \%(a d-), 24 \%(i n-),<1 \%(o b-), 21 \%(s u b-)$. Other linguistic effects are represented by a) case (the accusative has a noticeable positive effect on the presence of preposition, probability $=98 \%$ ), b) figurative usage of the verb, which has a positive effect too (probability $=97 \%$ ) and corroborates the hypothesis in Longrée (2005, p. 305) (see section 6.4.2.2), c) and animacy of the filler (there might be a small negative effect for animate fillers, probability $=72 \%$, although the result is not statistically reliable).

Finally, the literary ages all behave similarly, causing an increase in the probability of finding a preposition (the probability of 'construction' being 'preposition' is $91 \%$
for imperial and late republican), the exception being the Augustan era, which has a negative effect and probability of $92 \%$. This effect for 'era' is present even after factoring out the differences between the single authors throught the random effects: this means that the probability of finding a preposition does not only depend on the specific author, but also on the diachronic age.

### 7.4.4 First conclusions

We saw how hypothesis testing and regression modeling enabled us to discover relevant associations between the main variable of interest (presence vs. absence of preposition in the expression of the relatum of preverbs), specific chronological variables ('era' and 'author'), two morphological ones ('preverb' and 'case'), and two semantic ones ('animacy_relatum' and 'figurative'). In conclusion, the multivariate analysis described in section 7.4.3.2 showed the role played by other non-diachronic variables in the choice of the syntactic constructions for the PVs.

The question now arises as to whether the general diachronic trend detected in the first $\chi^{2}$ analysis is eventually confirmed or not. We will look for an answer by having recourse to a different technique and by extending the dataset.

### 7.5 Second investigation

So far, we have focussed on the alternation between construction 1 (CasePrev) and constructions 3 and 4 (Prep), which can be encoded as the opposition between presence and absence of the preposition in the realisation of the relatum of the preverb. This is because we modelled the alternation with a binary variable, which allowed us to make use of binomial distributions. These models gave us some information about the data, by showing the effects of specific linguistic variables. However, they impose strict constraints on the input (see section C.3.2), which are not necessarly met by our dataset.

Now we will make use of a different technique which does not impose strict constraints on the input data, while at the same time allowing us to focus on all four constructions for the morpho-syntactic expression of the relatum of a preverb (see section 6.4.1 on page 132):

1 (CasePrev): NP whose case is required by the preposition corresponding to the preverb;

2 (CaseNonPrev): NP whose case is different from the one required by the preposition corresponding to the preverb;

3 (PrepPrev): PP whose preposition corresponds to the preverb;
4 (PrepNoPrev): PP whose preposition does not correspond to the preverb.
We then collected the occurrences of all four constructions in the dataset and looked at the interaction between specific variables affecting them. This is the dataset for the second investigation, which consists of 353 observations.

### 7.5.1 The methodology

For this investigation, we opted for an exploratory approach (cf. page 140) using Simple Correspondence Analysis (CA; Benzécri 1973) and Multiple Correspondence

Analysis (MCA), a data-analysis technique widely used in social sciences and described in section C.4. The main reason for this methodological choice is that, unlike confirmatory data analysis, exploratory techniques do not require a priori theoretical hypotheses to be tested. Instead, they start by observing the dynamics taking place between the variables in the data; then, they propose a model for the account of the data evidence, as exemplified by the famous quote by Benzécri: "the model should follow the data. The data should not follow the model" (Benzécri, 1973).

Starting from a multi-dimensional representation of the dataset, CA and MCA produce low-dimensional plots that describe the essential structure of the data. The accuracy of this approximation is measured by the so-called percentage of explained inertia, which is the proportion of the variability present in the data that is captured by our analysis (see section C.4.3). Hence, we described our linguistic problem by using several sets of variables and comparing the percentages of inertia explained by the relative analyses: the higher the percentage of inertia, the best choice of variables. We focussed on two-dimensional plots, because they are easier to interpret and reproduce on paper.

### 7.5.2 Relevant variables

Although we adopted an exploratory approach, it should be pointed out that the particular choice of encoding of the data through specific variables is always bound to affect the analysis.

Since we are interested in detecting relevant variables affecting the choice of one construction over the other ones, we will illustrate two choices of variables, commenting on the results they produced.

### 7.5.2.1 Construction and author

We present a first exploration on the interaction between the two variables 'construction' and 'author' on the dataset described in section 7.4.1 and consisting of 237 PV tokens.

We then ran an MCA on the dataset classified by these two variables. Figure 7.6 displays the two-dimensional plot of the MCA, accounting for $88.6 \%$ of the total inertia. The details of the output are presented in section C.4.5.1.

In a CA (or MCA) plot, the points far from the origin are the most interesting ones, because they are associated with the least expected behaviour. Correlations between variables can be detected by looking at the angles between the corresponding vectors: small angles correspond to positive correlations, angles close to $180^{\circ}$ correspond to negative correlations, angles close to $90^{\circ}$ to no correlation.

From figure 7.6 it is possible to detect three clusters associating specific constructions to the single authors: construction 4 (PrepNoPrev) and Late Era authors (Jerome and Thomas), construction 3 (PrepPrev) and the early author (Plautus), and finally constructions 1 (CasePrev) and 2 (CaseNonPrev) and the Classical authors, with a particular preference for construction 2 (CaseNonPrev) by the poets.

Overall, we can state that the main diachronic result from the regression analysis described in section 7.4.3.2 is confirmed, since only the poets seem to prefer the constructions with no preposition: 1 (CasePrev) and 2 (CaseNonPrev). The prose writers' behaviour is close to expected. More specifically, this technique allowed us to detect a stronger preference of poets for the second construction (CaseNonPrev) than for the first one (CasePrev). This is in accordance with what we hypothesized in section 6.4.2.2. About the hypothesized higher frequency of construction 1 (CasePrev) in poetry and


Figure 7.6: Plot from MCA on the variables 'construction' and 'author'. The first axis accounts for $66.5 \%$ of the explained inertia, the second axis for $22.1 \%$. The labels 'construction.2CaseNonPrev' and 'Ovid' overlap.
post-Classical prose stated in section 6.4.2.1, the present analysis shows a positive association between this construction and poets; in addition, a positive association is detected for Petronius and Sallust too - which does not fit the hypothesis - although Petronius' behaviour is close to expected.

In addition, we were able to differentiate the behaviour of the Late Era authors (Jerome and Thomas) - who prefer the fourth construction (PrepNoPrev) - from the early author (Plautus), who prefers the third construction (PrepPrev). Again, this is in accordance with what was stated in section 6.4.2.3.

Finally, the analysis for Sallust, Petronius, Cicero and Caesar showed that these authors' behaviour is close to the expected one.

From a methodological point of view, this first exploration gives a preliminary picture of the distribution of the four constructions: unlike a one-dimensional representation, the two-dimensional plot allows us to display a wider range of possibilities for the position of the points corresponding to the variables of interest. For example, we can see that Thomas, while preferring construction 4 (PrepNoPrev), behaves closer to expected than Jerome; on the other hand, Caesar and Cicero prefer construction 3 (PrepPrev) but show also some preference for construction 1 (CasePrev). In section 7.5.2.2 we will see how this picture can be enriched.

### 7.5.2.2 More variables

The analysis given in section 7.5 .2 .1 is limited to the interaction between two variables: a chronological one ('author') and a morpho-syntactic one ('construction').

We took into account more variables in order to see how they affect the previous analysis:

- 'construction', values from 1 to 4: 'CasePrev', 'CaseNonPrev', 'PrepPrev', 'PrepNoPrev';
- 'era', a broad chronological classification of the authors of our dataset: 'early' (Plautus), 'classical' (Caesar, Cicero, Ovid, Petronius, Propertius, Sallust, Vergil) and 'late' (Jerome and Thomas);
- 'preverb';
- 'animacy': 'animate' if the preverb's relatum is animate, 'non.animate' otherwise;
- 'class', the semantic class of the verb: values 'motion', 'rest' and 'transport'.

The output of the analysis is presented in section C.4.5.2 and the two-dimensional plot of the MCA on this dataset is displayed in figure 7.7. The percentage of inertia explained by the first two axes of an MCA on the dataset described by all these variables is $48.0 \%{ }^{9}$ This figure is lower than that of the previous analysis (section 7.5.2.1), which is expected since the complexity of the model is higher. However, the main pattern detected in section 7.5.2.1 is kept constant, showing that the presence of these new variables does not change the basic structure of the data. In fact, the following associations are as before: between construction 4 (PrepNoPrev) and the Late Era authors (Jerome and Thomas), between construction 3 (PrepPrev) and the early author (Plautus), and between constructions 1 (CasePrev) and 2 (CaseNonPrev) and the classical authors. At the same time, the new analysis allowed us to obtain a geometrical picture of the interaction between a series of newly-introduced variables.

[^60]

Figure 7.7: Plot of the MCA for the second investigation on the interaction between 'construction', 'era', 'preverb', 'animacy', and 'class'. Percentages of explained inertia: 27.6 (first axis), 20.5 (second axis).

Figure 7.8 represents the four syntactic constructions under investigation as points in space. This representation of the dataset is the perspective under which we decided to examine the interrelations between the four values of the syntactic variable 'construction', as we will see now.

Animacy If the points of an MCA plot correspond to binary variables, we can inspect the vectors through the origin defined by joining the pairs of variables' values. In this case, 'animacy' is a binary variable, which defines a vector through the origin. On the upper right hand side, 'non.animate' (inanimate relata) is positively associated with the Late era and motion verbs. On the lower left hand side we find animate relata positively associated to both transport verbs with the Early Era, and to verbs of rest.

The correlation between inanimate relata and the Late Era authors might be connected to the genre of Thomas' and Jerome's works, theology and religious genre, respectively. In addition, the correlation between inanimate relata and construction 4 (PrepNoPrev) confirms what was argued by Longrée (2005) (see section 6.4.2.4 on page 137), but is in contrast with Rosén (1999, p. 138): "if we manage to disengage the semantic variables from the others, prepositional complements indeed come across as often being animate, and we may go along with the conclusion that the prepositional phrase is more 'concrete', if what it meant is its preferred occurrence in non-actantial positions".

Dramatic genre A higher frequency of construction 3 (PrepPrev) in the dramatic genre and the colloquial register is argued by López Moreda (1998, p. 964) (see page 136). In addition, the present analysis allowed us to acquire more information about this correlation in our corpora. Here we will consider the author from the dramatic genre recorded in the corpora (Plautus) and assume the generally recognised presence of colloquial features in his style (cf. e. g. Väänänen 1971; Poccetti et al. 1999; Rosén 1999). Then, we see that in our data Plautus' preference for construction 3 (PrepPrev) is not associated with all verb classes and preverbs, but it is especially related with verbs of transport and with the Source preverb $a b-$ 'from'.

Source/Goal asymmetry For what concerns the types of preverbs, the upper right hand side of the plot contains only Source preverbs ( $e x-$ and $d e-$ ), which generally tend to occur more often than expected with the constructions involving prepositions. On the left hand side of the plot we find Goal preverbs, concentrated on the upper half (apart from $a d$ - 'to', strongly preferred by Plautus); these preverbs are associated with the constructions involving the bare case.

If we accept this description of the complex picture we obtained, we can say that our result is generally in accordance with Imbert (2008)'s hypothesis illustrated in sections 6.4.2.3 and 6.4.2.2. Her hypothesis is based on the Source/Goal asymmetry, a topic which has been widely discussed in typology, syntax and psycholinguistics (cf. e. g. Ikegami 1987; Stefanowitsch and Rohde 2004; Lakusta and Landau 2005). Imbert (2008) employs the argument of the Source/Goal asymmetry in her explanation of the grammaticalization/lexicalization of Homeric and Old English preverbs. According to her proposal, Source preverbs have a higher tendency towards lexicalization - thus behaving as satellite preverbs - unlike Goal preverbs, which more frequently hold an adpositional role (relational preverbs). If we assume a correspondence between the satellite role of preverbs and their occurrence in constructions involving prepositions, and between the adpositional role and the occurrence in constructions involving the
bare case, we can say that this statement is overall confirmed by our data, although further specifications are in order, due to the complex picture of the phenomenon.

Methodological remarks The MCA analysis allowed us to proceed further in the investigation of the data by building on the results from the $\chi^{2}$ test and the regression model (section 7.4).

The $\chi^{2}$ test highlighted a diachronic trend, while the regression analysis showed that the detected trend is not only due to the author-specific differences (i.e. stylistic choices and genre), but to more general chronological groups. Moreover, the regression analysis identified other relevant variables affecting the syntactic constructions for the PVs. In particular, it showed that some preverbs ( $a b-$, ex- and de-) had a strong effect on the presence of the preposition. As to relating these preverbs with a diachronic variable, MCA detected a correlation between preverbs and chronological eras (Late for $d e$ - and $i n$-, and Early for $a b-$-). Therefore, the effect of these preverbs on the preference for the prepositional constructions can be interpreted diachronically. In addition, the distinction between the two constructions involving prepositions (PrepPrev and PrepNoPrev) supports the hypothesis of a diachronic trend: although both the Early Era author and the Late Era authors seem to prefer the use of preposition, they differ in the particular construction used (PrepPrev and PrepNoPrev, respectively).

### 7.6 Third investigation

So far, we have considered the morpho-syntactic realisations of the preverbs' relata. For the third investigation, we focussed on spatial complements occurring with PVs (e.g. advenio 'go from' + a goal complement), thus containing not only the preverbs' relata (e.g. advenio + a source complement). The reason for this is that we wanted to investigate the type of spatial expressions occurring with PVs and their relation with diachronic and semantic factors. Therefore, we used a larger dataset for the third investigation, to see if this allowed us to capture a higher degree of variability for the complex phenomenon under investigation.

In this section we describe the dataset for the third investigation (section 7.6.1) and illustrate the results of the analysis (section 7.6.2).

### 7.6.1 Dataset

We collected all instances of the selected PVs containing an expression of either the relatum of the preverb - constructions 1 (CasePrev), 2 (CaseNonPrev), 3 (PrepPrev), and 4 (PrepNoPrev) - or another spatial argument of the type Goal or Source; for the latter case we defined a positive occurrence of construction 5 (PrepOther) if it was a PP, and a positive occurrence of construction 6 (CaseOther) if it was an NP. These instances added up to 398.

Two examples of the fifth and the sixth construction are given by sentences (1) and (2).
(1) In villos abeunt vestes (Ov. Met. 1, 236)

Into hair:ACC.M.PL be transformed:IND.PRS.3PL garnment:NOM.F.PL
'The garnments change to hair'.
(2) abi rus (Plaut. Most. 8)
go:IMP.PRS.2SG countryside:ACC.SG.N
'go to the countryside!'.

### 7.6.2 Analysis

We are interested in finding whether the expression of a spatial complement of a PV depends on some specific chronological, morphological, lexical, syntactic and semantic factors. We then considered the variables listed below:

- 'preposition', distinguishing the prepositional constructions (PrepPrev, PrepNoPrev and PrepOther) from the NP ones (CasePrev, CaseNonPrev and CaseOther);
- 'era': Early, Classical, Late;
- 'complement': type of spatial complement, either G (goal) or S (source)
- 'animacy' of the lexical filler of the verb's complement: values 'animate' and 'non.animate'.

We performed an MCA to explore the relation between these variables in our data: the details of the output are reported in section C.4.6.

Figure 7.8 reports the plot from the analysis, which is a simplified 2-dimensional representation of the data, accounting for $82.6 \%$ of the total variance. The first axis accounts for most of the inertia (79.5\%). We can state that this axis is defined by the opposition between the presence and the absence of preposition with the PV: 'no.preposition' to the right and 'preposition' to the left, respectively associated with animate and inanimate fillers of the spatial complements. The opposition between Goal complements and Source complements is transversal with respect to the animacy opposition, showing a positive association between Source complements and presence preposition on the one hand, and between Goal complements and absence of preposition on the other.

The second axis constrasts the chronological distinctions by era: on top we find the Early Era, associated with both 'preposition' and 'no.preposition'; at the bottom we find the Classical and the Late eras, respectively associated with 'no.preposition' and with 'preposition'. In addition, we see that the Late era authors use a higher proportion of inanimate fillers than expected (which is compatible with the religious and philosophical topic of their works), the opposite holding for the previous authors.

Summing up, this investigation has confirmed a general statement about the presence of the preposition in the Late era and its absence in the Classical era, the exception being the Early Era author, who uses both. This is in contrast with the claim in (Brachet, 2000, pp. 66-67), which focusses on the constructions $a b \ldots a d \ldots d e$ - where a PV with the Source preverb de-occurs with a source PP with $a b$ and a Goal PP with ad. According to Brachet's proposal, this construction shows an instance of the old situation where de did not have an ablative sense; this construction precedes the later development represented by the construction $d e \ldots d e$-, where the verb prefixed with $d e-$ occurs with a PP introduced by $d e$.

In addition, the analysis allowed us to detect the correlation between Source complements and the use of the prepositions on the one hand, and between Goal complements and the use of bare cases on the other. This shows evidence for the Goal/Source asymmetry which we analysed in section 7.5.2.2.

Plautus The result for Plautus is in partial contrast with the previous analysis reported in section 7.5.2.2, where we detected a preference of Plautus for construction 3 (PrepPrev), a prepositional construction. Thus, if we consider the sixth construction (CaseOther) as well, Plautus suddenly appears to share a comparable preference for
constructions with and without preposition. This shows that his preference for prepositions is related with the preverb's relatum, while he prefers the bare case to express a different complement.

Plautus' status as a literary author is generally considered as coexisting with the presence of elements from the spoken language in his works: cf., among others, Fischer (1971) and Hofmann (1980). De Vivo (1994) reviews the main positions among the scholars, underlining the communicative nature of Plautus' language, aimed at adressing the large audience of his comedies; at the same time, De Vivo acknowledges the literary function of Plautus' style, and stresses the importance of sound effects in the performance on stage. This explains the frequent allitterations, parallelisms, repetitions and rhymes. As noted by Traina (1999, pp. 55 ff.), the prepositional construction PrepPrev can be considered among these figures of speech, because it involves some kind of repetition between the preverb and a preposition. Therefore, the particular position of Plautus in the detected diachronic trend should be evaluated in light of the features of this author's style, which help understand the high degree of variability in the constructions he uses.


Figure 7.8: Plot of the MCA for the third investigation. The first axis accounts for $79.5 \%$ of the explained inertia, while the second axis accounts for $3.2 \%$.

Differential Object Marking? After the previous general remarks, one could wonder if it is possible to detect instances of Differential Object Marking (DOM) in our Latin corpora. DOM is defined by Bossong (1983, p. 8) as "the subcategorization of direct objects, or, more precisely, of transitive patients, depending on the semantic properties of the object noun phrase". This phenomenon has been detected in a number of languages and the semantic properties involved are often related to either animacy or "referential features" such as definiteness. In the languages where DOM is based on animacy, transitive patients receive a differential (morphological) marking depending on their place along an animacy scale: for example, it is possible that objects denoting human beings are marked while those denoting animals are not, and so on from animals to inanimate objects. In Spanish, for example, if a direct object refers to a specific
person, it can be expressed with the preposition a 'to', as in busca a un médico '(s)he looks for a doctor' (Leonetti, 2004).

In the particular case we are studying, we compare constructions involving the presence vs. the absence of a preposition in the expression of a spatial complement of a Latin PV. From the previous analyses we found that the presence of a preposition generally increases with time (with exceptions) and that inanimate fillers are more frequent in later authors. We now ask if the rise of preposition-based constructions can be associated with a DOM feature in Latin: if this was the case, in earlier stages of the language non-prototypical fillers for spatial complements, such as animate ones, would receive the prepositional marking (advenio ad te 'I come towards you'), as opposed to the canonical expression involving the direct object (advenio Romam 'I go to Rome'). Such an investigation differs from the previous MCA study because it looks for an interaction between, for instance, animate fillers and presence of preposition in the classical poets within their overall preference for absence of preposition.

This idea leads us to phrase the following theoretical hypothesis:
Hypothesis 7.6.2.0.1. In the classical authors the combination of prepositional constructions - i.e. 3 (PrepPrev), 4 (PrepNoPrev) and 5 (PrepOther) - and the feature + ANIMATE is significantly higher than the combination of the same constructions and the feature -ANIMATE.

The results from a hypothesis testing based on Configural Frequency Analysis (von Eye, 1990) are reported on in section C.5. They show that the constructions without a preposition are associated with animate fillers in the classical authors, while the late era authors display the opposite behaviour. This confirms the general associations detected by the MCA analysis described earlier in this section; in addition, it shows that there is no evidence for a DOM hypothesis in our corpora, at least for what concerns the specific subset of Latin verbs consisting of the PVs we selected.

### 7.7 Summary and conclusions

The corpus-based quantitative analysis we carried out in this chapter allowed us to approach a complex linguistic phenomenon, for which different theoretical hypotheses have been proposed in the literature.

The general result confirms a diachronic trend which has been stated in historical linguistics studies and in grammaticalization theory, namely the increasing prevalence of prepositional constructions to express spatial complements of PVs. This fits the more general evolution from the synthetic features of Latin towards the analytical systems of Romance languages.

Our conclusion also fits previous theories on the development of preverbs which frame it as a grammaticalization/lexicalization process involving Path expressions (Imbert, 2008). This theory is based on the assumption that the presence of a preposition (PrepPrev or PrepNoPrev) could indicate a more advanced degree of lexicalization of preverbs, which is opposite to their original adpositional behaviour. This assumption focusses on syntactic evidence for a process which involves the semantics of the PV, thus acting at the syntax-semantics interface.

In addition to this broad tendency, the multivariate techniques we used (regression modeling, correspondence analysis and configural frequency analyis) made it possible to conduct advanced investigations, which led to more detailed conclusions. We further specified the general trend by recognising that the late era authors of our corpus prefer
the use of prepositions that are different from the ones corresponding to the preverb of the PV (PrepNoPrev). On the other hand, the Early author (Plautus) prefers the same preposition as the preverb (PrepPrev) when expressing the preverb's relatum, and a bare noun phrase for other complements. Finally, we found some evidence supporting the Goal/Source asymmetry, associating the presence of the preposition with Source complements and its absence with Goal complements.

These results refer to single authors as representatives of chronological eras. This is a tricky question, given the fact that we cannot disregard stylistic factors affecting the authors' choices, as well as metric constraints in poetic texts, not to mention the peculiarities of the dramatic genre. The case of Plautus - the only Early Era author in our corpus - is particularly indicative of this. As we discuss in section 7.6.2, his choice of which constructions to use is also driven by stylistic factors, which make him introduce expressive repetitions and sound effects. In addition, we have found that poets in our corpora tend to avoid the use of prepositions in the realisation of the preverbs' relata, while Plautus shows an opposite behaviour, although with some restrictions. This result could be viewed in light of Roman Jakobson's studies (Jacobson, 1966) according to which the poetic function focusses on the linguistic form. In our case, the emphasis on the preverb within the PV (and therefore on the constructions with the bare case) is more expected in poetry and less expected in spoken language and our findings seem to confirm these expectations.

### 7.7.1 Future work

The next step in this reseach could involve accounting for word order features in collecting the verbs' arguments, as well as a measure of the distance between the different constructions used within the same text, in order to detect additional relevant variables. In addition, the frequency of a construction with a particular verb could be compared with the overall frequency of the verb, since it has been hypothesized that some constructions decrease in frequency together with a decrease in the frequency of verbs (Longrée, 2005, p. 303). Another variable to take into account is the type of clause the PVs occurs in (main or subordinate) and the distance between verb and complement, which is said to influence the prepositional use (Longrée, 2005, p. 305).

We believe that the methodological contribution of this preliminary case study will serve as a basis for further investigations, which may make use of a larger corpus. For example, it would be good if other archaic poets were included, as well as prose writers from the Augustan era and a wider range of genres from the later eras. The availability of a larger corpus would also allow us to statistically test the validity of the results found, proceeding towards their generalization to a larger portion of the Latin language.

## Chapter 8

## Conclusions

### 8.1 Introduction

In this chapter we comment on the main contributions of the present thesis. First, we summarise our most relevant findings and achievements in the light of the goals stated in chapter 1 (section 8.2). Second, we suggest future directions of research (section 8.3). Finally, we give some general conclusions (section 8.4).

### 8.2 Aims and contributions

Chapter 1 set the goals for the present research: provide a new lexical resource for the study of Latin verbs and propose a computational methodology for Latin linguistics.

This section stresses our contribution in relation to the research aims, both from the perspective of the new resource created (section 8.2.1) and from a methodological point of view (section 8.2.2).

### 8.2.1 New resource

The new resource we have created is a corpus-based subcategorization (SC) lexicon, described in chapter 4 ; we have also enriched the lexicon with selectional preference (SP) information, as explained in chapter 5.

The lexicon is an advance in the state of the art of computational resources for Latin, which is a low-resource language. The resources currently available for Latin include not only digitalised editions of texts, but also morphologically and syntactically (but not semantically) annotated corpora; however, their size is still small compared with that of corpora for some modern languages. Our lexicon places itself in this line of research by offering a further step in the direction of creating advanced resources and making them available in a way which provides the scientific community with rich linguistic data.

More specifically, the lexicon contains information on the verbal arguments occurring in the corpora we used (LDT and IT-TB): in addition to SC patterns, it includes information on verbal morphology, as well as morphological and lexical features of the arguments. We give a "usage-based" definition of the SC structures included in the lexicon, which contrasts with the intuition-based approach followed in traditional Latin dictionaries; our approach was inspired by the research on acquisition of SC lexicons
in computational linguistics. While traditional dictionaries and lexicons record the patterns which, according to the lexicographers, are considered to pertain to verbs' valency, the SC structures of our lexicon are the patterns actually occurring in the reference corpora. This perspective allows for further corpus-based synchronic and diachronic investigations on verbal argument structure which account for language use.

From the point of view of its usability, the lexicon can be easily searched by linguists who do not program thanks to the graphical interface currently under development. The use of the lexicon ranges from searches aimed at simply browsing the corpora looking for instances of verbal arguments, to the extraction of more comprehensive data to be used in quantitative studies.

### 8.2.2 New methods

The methodological contribution of this thesis regards the computational and quantitative approach to the study of Latin verbs that we propose. This kind of approach obviously requires a great deal of technical expertise, while at the same time demanding solid understanding of the theoretical issues involved.

First, our SP acquisition system is an example of automatic semantic processing of Latin; it also shows how the low-resource status of Latin can be overcome by techniques which build on corpus annotation and go beyond it. Since there are no semantically annotated corpora for Latin, our system exploits both existing resources for Latin (treebanks and Latin WordNet) and computational models developed for extant languages (Alishahi, 2008). In this process, we tailored Alishahi's model to the features of the Latin corpora we used: this adaptation was challenging because it required solutions to several problems, among which the small size and the low coverage of the resources available were the most severe.

Moreover, the evaluation phase gave us the chance to compare the performance of a supervised system which relies on WordNet with an unsupervised one based on a distributional measure of similarity calculated on a much larger automatically parsed corpus. The results show that the former system performs better, which suggests that the supervised procedures might currently be the best choice, given the less-resourced status of Latin. However, more research is needed to investigate whether other distributional models, paired with large corpora, can provide better results: this would pave the way for a more unsupervised approach to Latin computational linguistics.

Second, a further methodological contribution of our work regards the case study on Latin preverbs, reported on in chapter 7. Ours is the first study on Latin preverbs which at the same time is quantitative and uses rich data from freely available corpora. This gave us the chance to demonstrate the potential of statistical techniques for diachronic linguistics. At the same time, our systematic use of corpora as the basis for the analysis is in accordance with usage-based approaches in (historical) linguistics.

The results of our study show that the argument structure of Latin prefixed verbs is a complex phenomenon: a number of different morphological, lexical and semantic factors interact with diachronic variables showing a very rich picture. The techniques we employed are a very good way to detect and quantify variation in linguistic data, and thus identify patterns in highly complex data. Following a view shared by several scholars in different frameworks (see e. g. Milroy 1993; Denison 2003; Lightfoot 2006), we consider variation as the key for language change. Therefore, thanks to these statistical techniques able to highlight variation, we detected a general diachronic trend which confirms a tendency towards the use of prepositional phrases with prefixed verbs even when the bare case occurred in the classical era. This change can be connected with a
lexicalization process and framed within the shift that took place from satellite-framed (Latin) to verb-framed (Romance languages) according to Talmy's typology of space.

The statistical analyses also show that for some authors in our corpora there is so little data that it is not possible to give a reliable estimation. This is very valuable because it offers an empirical answer to the question of what should be included in our corpus in order to achieve reliable results: by quantifying the level of confidence of our estimates, these techniques are able to accurately identify possible problems in the composition of the corpus.

While describing the quantitative analysis, we also had the chance to apply a methodological procedure which is common in data analysis research to Latin linguistics. According to this procedure, the phases of collection, analysis and interpretation of the data are kept distinct, in contrast to the traditional and usual procedure followed in Latin linguistics. The complexity of the phenomenon studied is matched by the richness of the input data collected from the computational resource we built (the lexicon). This complexity is handled by specific statistical techniques (such as multivariate techniques) that account simultaneously for the contribution of the different factors; this is not possible with simpler approaches based on e.g. percentages or $\chi^{2}$ tests. The effect of this is that the linguistic interpretation rests on solid empirical grounds.

### 8.3 Future research

The research reported here covers a range of linguistic fields, each of which can be further developed.

First, the method for extracting the SC lexicon is fully automatic, hence it can be adapted to other Latin treebanks, both existing ones (such as the PROIEL treebank) and future ones. Furthermore, for the same reason the lexicon can of course be extended as the input treebanks grow in size. Additionally, more information could be included in the lexicon, such as verb senses, the distinction between optional and obligatory arguments of verbs, SC frame alternations. This can be achieved by drawing upon previous research in computational linguistics, where these tasks have been faced by a number of supervised and unsupervised systems.

Second, the SP acquisition system can be improved by including a Word Sense Disambiguation module that assigns different weights to the different word senses contributing to the definition of semantic properties, depending on their frequency. In addition, as we said in section 8.2.2, for the evaluation we performed a comparison between the WN version of the system and the distributional one. This comparison could be further developed in order to find the best distributional model for SPs of Latin verbs. This result would help the performance of the system and, at the same time, provide new insights on the semantics of Latin verbs and on their syntax-semantics interface.

Third, the results obtained on Latin spatial preverbs have a number of possible future developments. The methodology we proposed, based on multivariate statistics, can be applied to the study of other important variables potentially affecting the realisation of verbal arguments of prefixed verbs. In addition, the range of authors analysed could be increased thanks to the availability of new annotated corpora. In particular, if this reduced the unique associations between chronological eras and genres, it would be possible to assign the trend we have detected to either stylistic or diachronic effects.

### 8.4 General conclusions

This thesis is intrinsicly cross-disciplinary, because it involves considerations and methods from computational linguistics and historical linguistics. Our work will have hopefully illustrated what can be achieved when traditional barriers between subdisciplines are trascended.

Ideally, as this thesis shows, such work should be carried out by researchers with a non superficial command of both subfields, although this might sometimes be difficult to achieve in practice. In lieu of the widespread availability of such cross-disciplinary background, we propose that both historical linguists and computational linguists have a responsability to acquaint themselves with each other's fields and together explore how the study of linguistic phenomena can be furthered. We also hope that computational and historical linguists alike will be inspired by this work and seek to establish fruitful collaborations in the future.

Finally, we have demonstrated that computational linguistics has more to offer to historical linguists than a set of tools; computational linguistics provides a wide selection of flexible models and techniques that in themselves can provide new insights about historical processes. As such, this constitutes a new perspective in the study of linguistic phenomena, in the spirit of the proposals set forth in Geeraerts (2006). We believe that research can only progress through such an approach. Far from substituting each other's roles, historical and computational linguists can enter into a productive simbiosis.

## Appendices

## Appendix A

## Background to chapter 4

This appendix provides the technical details of the SC extraction system we developed for building the lexicon described in chapter 4.

Section A. 1 illustrates the format of the two treebanks we used (LDT and IT-TB), while section A. 2 explains the database queries we designed. Section A. 3 describes the different types of SC structures we included in the lexicon and finally section A. 4 illustrates how to extract frequency information from the lexicon.

## A. 1 Annotation format of the treebanks

The LDT encodes dependency structures into XML documents. Table A. 1 records the XML code for sentence (2) (page 57) from the LDT. Each sentence corresponds to a XML tag called 'sentence' and its unique identifier is given in the form of the attribute 'id', together with other attributes related to the specific work and author in question. Each token of the sentence (except punctuation marks at the end of the sentence, which are not recorded in the LDT) corresponds to a tag called 'word', whose attributes are: ${ }^{1}$

1 'id': position (rank) of the token in the sentence (1),
2 'form': inflectional form (quem),
3 'lemma': lemma (quis1), ${ }^{2}$

[^61]<sentence $\mathrm{id}=$ "390" document_id="Perseus:text:1999.02.0010">
<word id=" 1 " form="quem" lemma="quis1" postag="p-s---ma-" head=" 3 " relation="ATR" />
<word $\mathrm{id}=" 2$ " form="ad" lemma="ad1" postag="r- . . . ." head=" 6 " relation="AuxP" />
<word id=" 3 " form="finem" lemma="finis1" postag="n-s- - -ma-" head=" 2 " relation="OBJ" /> <word id=" 4 " form="sese" lemma="sui1" postag="p-s--fa-" head=" 6 " relation="OBJ" $/>$
<word id $=" 5$ " form="effrenata" lemma="effreno1" postag="t-srppfn-" head=" 7 " relation="ATR" $/>$
<word id=" 6 " form="iactabit" lemma="jacto1" postag="v3sfia---" head $=$ " 0 " relation="PRED" / > <word id=" 7 " form="audacia" lemma="audacia1" postag="n-s---fn-" head="6" relation="SBJ" / > $</$ sentence $>$

Table A.1: XML code for sentence (2)(page 57) in the LDT.

```
<s id="001.1SN.DS-4QU2.AR2-AG-1.2-6.3-4" >
\(<\mathrm{f}>\) praedicatum \(<\mathrm{l}>\) praedicatum \(<\mathrm{t}>11 \mathrm{~B}-\mathrm{-x}-\mathrm{A} 3--1<\mathrm{A}>\mathrm{Sb}<\mathrm{r}>1<\mathrm{g}>3\)
\(<\mathrm{f}>\) enim \(<\mathrm{l}>\) enim \(<\mathrm{t}>4\)-O------ \(<\mathrm{A}>\) Coord \(<\mathrm{r}>2<\mathrm{g}>0\)
\(<\mathrm{f}>\) habet \(<\mathrm{l}>\) habeo \(<\mathrm{t}>3\)-KA1- \(-6--<\) A \(>\) Pred_Co \(<\mathrm{r}>3<\mathrm{g}>2\)
\(<\mathrm{f}>\) rationem \(<\mathrm{l}>\) ratio \(<\mathrm{t}>11 \mathrm{C}--\mathrm{D} 2---\mathrm{A}>\mathrm{Obj}<\mathrm{r}>4<\mathrm{g}>3\)
\(<\mathrm{f}>\) formae \(<\mathrm{l}>\) forma \(<\mathrm{t}>11\) A- - -B2- \(-1<\mathrm{A}>\) Atr \(<\mathrm{r}>5<\mathrm{g}>4\)
\(<\mathrm{d}>.<\mathrm{l}>.<\mathrm{t}>-\mathrm{-}-\mathrm{-}-\mathrm{-}-<\mathrm{A}>\mathrm{AuxK}<\mathrm{r}>6<\mathrm{g}>0\)
```

Table A.2: CSTS code for sentence (3) in the IT-TB.

4 'postag': morphological features ( $p-s-$ - $-m a-$ ),
5 'head': rank of the token's dependency head (3),
6 'relation': dependency relation that links the token to its head (ATR).
If a word depends on the sentence node, its head is 0 . Among the morphological features for quem in table A.1, ' p ' stands for ' $\mathrm{PoS}=\mathrm{pronoun}$ ', 's' for 'number=singular', ' $m$ ' for 'gender $=$ masculine', and ' $a$ ' for 'case=accusative'. ${ }^{3}$

For encoding its annotation, the IT-TB uses the CSTS (Czech Sentence Tree) format. Table A. 2 presents the CSTS version of the tree in figure 4.3 (page 58). In the IT-TB each sentence corresponds to a tag 's', whose attribute is the unique identifier for the sentence ('id'). In addition, each token corresponds to a line containing the following six tags: ${ }^{4}$

1 ' f ': form (praedicatum),
2 'l': lemma (praedicatum),
3 ' $t$ ': morphological tags (11B--A3--1),
4 'A': functional label ( $S b$ ),
5 'r': rank of the token in the sentence (1),
6 ' g ': rank of its head in the dependency tree (3).

Punctuation marks are tagged as 'd', followed by the tags just listed. Regarding the morphological tags for praedicatum in table A. 2 , ' 1 ' in the first position stands for 'nominal flexional type' (for nouns, adjectives and pronominals), ' 1 ' in the second position stands for 'positive degree', 'B' for 'second declension', 'A' for 'singular nominative', '3' for 'neuter', and the last ' 1 ' for 'baseform graphical variation'. The full tagset is illustrated in detail at http://gircse.marginalia.it/\~passarotti/tagset/IT\ tagset.pdf.

[^62]
## A. 2 Database queries for subcategorization extraction

In section 4.5 .3 we briefly outlined the procedure we developed to extract SC structures from our two treebanks. Here, we go into the actual queries we designed.

In order to detect every occurrence of the arguments of verbs in the two treebanks, we loaded both corpora into two database tables called 'ldt' and 'it', respectively.

Figure 4.11 is a screenshot of the database interface with the entry of sentence (2) from the 'ldt' table (see figure 4.2 on page 57 ; for the corresponding XML entry, see table A.1). Each row (database record) represents a word of the sentence and each column (database field) corresponds to one of the XML attributes explained on page 171. The fields for the 'ldt' table are: 'ID', 'form', 'lemma', 'pos', 'person', 'number', 'tense', 'mood', 'voice', 'gender', 'case', 'degree', 'relation', 'rank', 'head', 'sentence'; we also inserted 'author' as an additional field.

Figure A. 1 shows a screenshot of the database interface with the entry of sentence (3) from the 'it' table (see figure 4.3, page 58; for the CSTS entry see table A.2, page 172). Compared with the original CSTS file, we identified each word with its unique 'ID' and associated the sentence ID with the new field 'sentence'. In addition, we splitted the 't' attribute into the morphological features 'pos' (PoS), 'degree_nom' (degree of nominals), ${ }^{5}$ 'fl_cat' (flexional category, like declension or conjugation), 'mood', 'tense', 'degree_part' (degree of participials), ${ }^{6}$ 'case', 'gen_num' (gender/number/person), 'comp' (composition for enclytics and compounds), 'var' (formal variation), ${ }^{7}$ and 'grap_var' (graphical variation). ${ }^{8}$ Finally, 'relation' contains the functional label (tag ' f ' in the CSTS code), 'rank' its position in the sentence (tag ' r '), 'head' its syntactic head (tag ' $g$ ').


Figure A.1: Sentence (3) (page 58) of the IT-TB as it is loaded into the database.

While extracting the SC lexicon, we searched the databases through MySQL queries. In the next sections we will illustrate the procedure that led to these queries, giving

[^63]examples from the corpora.

## A.2.1 Searching for verbal tokens

We first searched for all occurrences of verbs in the treebank: this search is based on a morphological criterion: the PoS must be a verbal predicate, annotated as ' $v$ ' (verbs) and ' $t$ ' (participles) in the LDT and as ' 2 ' (participials) and ' 3 ' (verbs) in the IT-TB. These are the queries that define the 'verbs' table for each database:

Query A.2.1. SELECT * FROM Idt WHERE Idt.pos=" $v$ " OR Idt.pos=" " $t$ "; SELECT * FROM it WHERE it.pos=" 2 " OR it.pos=" 3 ";

Hence, we aim at extracting verbs occurring at any position in the tree, not only verbs syntactically tagged as 'Pred', since each such verbal occurrence constitutes a valency instance that must be recorded in the SC lexicon. It should be noted that multiword verbal expressions are not detected by query A.2.1. Thus, in sentence (4) on page 59 the verbal lemma fio only (and not the whole expression certior fio) is detected and accounted for in the lexicon. In addition, special constructions such as those involving the impersonal usages of paeniteo and pudeo are recorded in the lexicon according to their annotation in the treebanks. For example, in the sentence Pudet haec opprobria nobis et dici potuisse et non potuisse refelli (Ovid, Met., I, 758-759; 'I am ashamed that these reproaches are objected to me and cannot be answered'), pudet is lemmatized as pudeo, opprobria and the two occurrences of potuisse are tagged as 'Sb_Co' depending on pudet and nobis is annotated as its subject: therefore, the SC class of pudeo in this sentence is ' $\mathrm{Obj}, \mathrm{Sb}$ '.

## A.2.2 Searching for verbal arguments: direct head-argument dependencies

As we were looking for verbal arguments in the treebanks, we identified four functional labels which mark them (see page 65 ff . for a description of these tags):

- 'Sb': subject;
- 'Obj': direct and indirect object;
- 'Pnom': predicate nominal;
- 'OComp': object complement.

The set of verbal arguments is extracted by looking at all database records satisfying the following search query:

Query A.2.2. SELECT * FROM db WHERE relation IN('Sb', 'Obj', 'Pnom', 'OComp');
This pseudo-query searches the database 'db' (either 'ldt' or 'it') for all records whose 'relation' field has values in the list $\{$ ' Sb ', 'Obj', 'Pnom', 'OComp'\}. We called the nodes obtained this way as target nodes, and defined the 'target' table as the table whose rows consist of these nodes.

According to the annotation style of the treebanks, the arguments of a verb depend on the verbal head, as illustrated in figure A.2. For example, the subject praedicatum and the object rationem in figure 4.3 (page 58) are dependent on the verb habet. This dependency is encoded in the 'head' field of the 'ldt' and 'it' tables. As can be seen in

Verb<br>(VERBAL NODE)<br>Sb, Obj, Pnom, or Comp<br>(ARGUMENT NODE)

Figure A.2: A 'direct' dependency between a verbal head (annotated with any syntactic tag) and an argument node (tagged as 'Sb', 'Obj', 'Pnom', or 'OComp').
figure A.1, the head of the words praedicatum and rationem is ' 3 ', which is the rank of the verb habet. The dependency criterion to detect all verbal arguments of a verb ' v ' (whose rank is 'rv') is:

Query A.2.3. SEARCH FOR ALL argument RECORDS IN target SUCH THAT
head $($ argument $)=r v$ and sentence(argument) $=$ sentence $(v)$;
In this pseudo-code, we represented the query that searches for all records in the 'target' table headed by ' $v$ ' and occurring in the same sentence as ' $v$ '. Queries similar to query A. 2.3 were designed to create the 'tree' table, that is the table containing all word-parent pairs, where the word node and the parent node (identified by their IDs) are all nodes connected by a tree arc. Table A. 3 shows the portion of the 'tree' table for all the word-parent pairs in sentence (3). ${ }^{9}$

The subset of 'tree' containing all word-parent pairs where the word belongs to the 'target' table and the parent to the 'verbs' table was called 'path'. 'Path' records the depth of the relation between the word and its parent in the third column: this depth is 1 for the direct relations analysed so far. Table A. 4 shows a the portion of the 'path' table for all the target-verb pairs in sentence (3) (page 58; see also figure 4.3 on page 58).

| word | parent |
| :---: | :---: |
| praedicatum | habet |
| habet | enim |
| rationem | habet |
| formae | rationem |

Table A.3: 'Tree' table for sentence (3) from the IT-TB.

|  | verb | depth |
| :---: | :---: | ---: |
| praedicatum | habet | 1 |
| rationem | habet | 1 |

Table A.4: 'Path' table for sentence (3) from the IT-TB.

[^64]
## A.2.3 Searching for verbal arguments. Indirect head-argument dependencies

The description above treats the simple cases where the dependency between the verbal head and the argument node is direct, as in figure 4.3 (page 58).

However, due to the annotation style, more complex cases arise when "bridge" nodes intervene between the verbal headword and the its arguments. These intermediate nodes can be: ${ }^{10}$

- 'AuxP': prepositions;
- 'AuxC': conjunctions;
- 'Coord': coordinating elements;
- 'Apos': apposing elements.
'AuxP' nodes usually consist of prepositions which link a verb with an NP; this NP can be either an adjunct (tag 'Adv', like in-7 hibernis-8 in figure 4.4, page 60 ), or an argument. For the SC extraction, we are interested in the latter case. Let us consider the PP ad-16 eum-17 in figure 4.4 (repeated in figure A.3): here, the argument eum is attached to the verb adferebatur-19 via the preposition $a d$, as shown in figure A.3.
adferebatur
Pred_Co
(VERBAL NODE)
ad
AuxP
(INTERMEDIATE NODE)
।
eum
Obj
(ARGUMENT NODE)

Figure A.3: An indirect dependency between a verbal head and its argument via an 'AuxP' node. This example is taken from sentence (4).

Similarly, 'AuxC' nodes consist of conjunctions usually linking a verb with the predicate of a subordinate clause; this subordinate clause can either act as an adjunct (tag 'Adv'; e. g. node venisset-9 in figure 4.9, page 64), or as an argument. For the SC extraction we aim at retrieving the latter cases. An example is proficisceretur-9, predicate of the subject clause ad eos proficisceretur and depending on the predicate dubitandum-3 (esse) via the conjunction quin-6 in figure 4.7 (page 63). From the examples above we conclude that the search queries must include 'AuxP' and 'AuxC' in the list of candidate intermediate nodes. We called this list 'inpath1':

Query A.2.4. SELECT * FROM db WHERE relation IN('AuxC','AuxP');
In this pseudo-code we represented a query that searches the database 'db' (either 'ldt' or 'it') for all records whose 'relation' field has values in the list \{'AuxC', 'AuxP'\}.

[^65]'Inpath1' combines with 'target1', the list of arguments that allow 'AuxP' and 'AuxC' to be intermediate nodes between the argument nodes and the verb, defined by the following query:

Query A.2.5. SELECT * FROM db WHERE relation IN('Sb', 'Obj', ‘Pnom', 'OComp');
This pseudo-database query searches the database 'db' (either 'ldt' or 'it') for all records whose 'relation' field has values in the list \{'Sb', 'Obj', 'Pnom', 'OComp'\}.
'Coord' (see page 67 ff .) annotates coordinating elements. When a verb has two or more coordinated arguments, a 'Coord' node intervenes between the verb and its argument nodes. Analogous cases can be found with the 'Apos' node. By way of example of coordinated arguments, consider the objects of the verb dat-1 in figures 4.6 (page 62 ) and A. 4 (page 177): senonibus-3 and Gallis-6. ${ }^{11}$ These nodes depend on the verb dat-1 via the coordinating node que-4.


Figure A.4: An indirect dependency between a verbal head and its arguments via a 'Coord' node. This example is taken from sentence (4).

More complex cases are possible, where both types of intermediate nodes ('AuxP'/ 'AuxC' and 'Coord'/ 'Apos') may occur. In addition to objects in coordination (or apposition) depending on a verb, there can be multiple coordinated (or apposed) PPs (introduced by 'AuxP') or subordinate clauses (introduced by 'AuxC'). In these cases, the affix '_Co' (or '_Ap') is appended to the children of the 'AuxP' (or 'AuxC'), since 'AuxP_Co' and 'AuxC_Co' are not accepted tags. As an illustration of this, see sentence (1) below, whose tree is represented in figure A.6.

```
(1) sed nomen rei imponitur a quidditate vel
    but name:NOM.N.SG thing:GEN.F.SG assign:IND.PRS.3SG.PASS by quidditas:ABL.F.SG or
    forma (Th., Sent. P. L., I, D. 25, Q. 1, Art. 4, 6-1.7-7
    form:ABL.F.SG
    'Conversely the name reality is assigned by quidditas or by the form'
```

The verb impono here occurs with two coordinated indirect objects, quidditate and forma. In such cases, both 'AuxP' and 'Coord' are intemediate nodes that need to be extracted together with the coordinated arguments.

[^66]

Figure A.5: Coordinated PPs with one 'AuxP' node.


Figure A.6: Dependency tree of sentence (1) from the IT-TB.

Arguments can also appear with prepositions in coordination depending on a verb (figure A.7). For example, the following sentence shows two coordinated PPs: a popularibus coniurationis and a multitudine conducta.


Figure A.7: Coordinated PPs with two 'AuxP' nodes.
(2) itaque censuit
pecunias eorum
publicandas,
and so deliberate:IND.PF.3SG money:ACC.F.PL they:GEN.M.PL confiscate:GERUNDIVE.ACC.F.PL,
ipsos per municipia in custodiis habendos, videlicet they:ACC.M.PL in town:ACC.N.PL under guard:DAT.F.PL hold:GERUNDIVE.ACC.M.PL, of course timens, ne, si Romae sint, aut a popularibus fear:PTCP.PRS.NOM.M.SG, that, if Rome:LOC.F.SG be:SBJV.PRS.3PL, either by popular:ABL.M.PL coniurationis aut a multitudine conducta per conspiracy:GEN.F.SG or by crowd:ABL.F.SG guide:PTCP.PF.ABL.F.SG through $\operatorname{vim}$ eripiantur. (Sall., Cat., 52)
violence:ACC.F.SG rescue:SBJV.PRS.3PL
'And so he deliberates that their money should be confiscated by the state, and that they should be held under guard in various free-towns, no doubt fearing that if they are at Rome they might be forcibly rescued by their co-conspirators or by a hired crowd.'

Therefore, a specific search query must be designed to include 'Coord' and 'Apos' in the list of candidate intermediate nodes: we called this list 'inpath2'.

Query A.2.6. SELECT * FROM db WHERE relation IN('AuxC', 'AuxP', 'Coord\%', 'Apos\%');
In this pseudo-code we represented a query that searches the database 'db' (either 'ldt' or 'it') for all records whose 'relation' field has values in the list \{'AuxC', 'AuxP', , 'Coord_Co', 'Coord_Ap', 'Apos_Co', 'Apos_Ap'\}.
'Inpath2' combines with 'target2', the list of arguments that allow 'AuxP', 'AuxC', 'Coord_Co','Coord_Ap', 'Apos_Co' and 'Apos_Ap' to be intermediate nodes between the argument nodes and the verb, defined by the following query:

Query A.2.7. SELECT * FROM db WHERE relation IN('Sb\%', 'Obj\%', 'Pnom\%', 'OComp\%');
This pseudo-query searches the database 'db' (either 'ldt' or 'it') for all records whose 'relation' field has values in the list $\{$ ' $\mathrm{Sb} \%$ ', 'Obj\%', 'Pnom\%', 'OComp\%'\}; the ' $\%$ ' sign indicates that the elements of this list are allowed to append an affix ('_Co' or '_Ap') to the tags 'Sb', 'Obj', 'Pnom', and 'OComp'.

Table A. 5 shows the tables 'inpath1', 'target1', 'inpath2', and 'target2' (see pages 176 ff .) for sentence (1). ${ }^{12}$ In such cases of indirect relations, the 'path' table contains

| table | content |
| :--- | :--- |
| inpath1 | $a($ auxP $)$ |
| target1 | nomen (Sb) |
| inpath2 | $a$ (AuxP), vel (Coord) |
| target2 | quidditate (Obj_Co), forma (Obj_Co) |

Table A.5: 'Inpath1', 'target1', 'inpath2', and 'target2' tables for sentence (1) from the IT-TB.
the target-parent pairs where the target node belongs to the 'target1' or 'target2' table and the parent is either a verb node (belonging to the 'verbs' table) or an intermediate node between the verb and an argument node (belonging to 'inpath1' or 'inpath2'). Consequently, 'path' may now contain numbers greater than 1 in the 'depth' column, because of the indirect verb-argument relations.

Table A. 6 shows the portion of the 'path' table for all target-parent pairs in sentence (1). ${ }^{13}$ The 'depth' column is 1 when the row contains a target-verb pair (quidditateimponitur, forma-imponitur, and nomen-imponitur), 2 when it contains a target-parent

[^67]| target | parent | depth |
| :---: | :---: | ---: |
| quidditate | imponitur | 1 |
| forma | imponitur | 1 |
| nomen | imponitur | 1 |
| quidditate | $a$ | 2 |
| forma | a | 2 |
| quidditate | vel | 3 |
| forma | vel | 3 |

Table A.6: 'Path' table for sentence (1) from the IT-TB.
pair where the parent node is an intermediate node directly depending on the verb (quidditate-a and forma-a); finally, 'depth' is 3 for target-parent pairs where the parent node depends on the verb via another intermediate node (quidditate-vel and forma-vel). Note that 'Pnom' and 'OComp' are considered as target nodes and cannot be part of a path. For this reason, the argument structure of the verb fio in sentence (4) (see figure 4.4) includes only the 'Pnom' argument certior-24 and not the accusative + infinitive clause Belgas ...contra populum Romanum coniurare obsidesque inter se dare.

## A.2.4 Searching for verbal arguments. Some difficult cases

The way coordination and apposition are handled in the treebanks causes arguments of coordinated or apposed verbs to appear as depending on the coordinating or apposing node rather than on the verbal nodes. This case of "external" argument nodes is illustrated in figure A.8. For example, in figure 4.4 (page 60), the subject Belgas-27 is


Figure A.8: "External" argument nodes.
syntactically associated to both verbs coniurare-39 and dare-44, but it is attached to the coordinating node que-40 in the tree.

As the coordinated (or apposed) verbal nodes can have any syntactic tag in such cases, the only way to distinguish the external argument node from the verb pair is to look at its syntactic tag, which should have no affixes. Such cases are included in the SC lexicon through a special query. It can also happen that the arguments are coordinated (or apposed) and they both syntactically depend on a verb pair.

A more difficult case is when both the verb pair and the argument pair have the same syntactic tags, as in the following sentence, represented in figure A.9:

It is surprising that John and Philip read and write.


Figure A.9: Dependency tree of sentence (3) from the IT-TB.

In this case, we have a verb pair (read and write) annotated as 'Sb_Co' (as they are the predicates of a subject clause depending on the main clause it is interesting), and an argument pair (John and Philip), again annotated as 'Sb_Co', as illustrated in figure A.9. As this is an instance of a potentially ambiguous case, we decided to isolate the sentences that satisfy this condition and treat them separately and manually in the SC lexicon.

## A. 3 Subcategorization frames and classes

In section 4.6.1 we introduced the SC structures we defined: SC frames (SCFs; recording the order of the elements in the pattern) and SC classes (SCCs).

In order to account for the syntactic roles of the arguments, both the SCFs and the SCCs show the arguments' functional labels, contained in the 'relation' field of the databases 'ldt' and 'it'. SCFs and SCCs are built from the 'path' table (see page 174 ff .) by combining all the nodes depending on a verbal node, including direct and indirect dependencies (see page 174 ff .). In the latter case, the path is recorded as well.

For example, the two rows of table A. 4 (page 175) correspond to the two arguments of habeo in sentence (3). If we use their IDs to retrieve from the original tables 'ldt' and 'it' their functional labels ('Sb' and 'Obj' respectively) and their ranks (1 and 4) compared to the verb's rank (3) (see also table A.2), we are able to claim that in sentence (3) the verb habeo is preceded by a subject (praedicatum) and followed by an object (rationem). The SCSs precisely record this type of information. So, the SCF and the SCC for the verb habeo in figure 4.3 are respectively:

$$
\mathrm{SCF}: \mathrm{Sb}+\mathrm{V}+\mathrm{Obj}
$$

SCC: Obj, Sb

We will now describe more in detail the different SCSs making up the lexicon's entries.

These structures are listed in table A.7, ranging from the most specific to the most general one. SC frames and SC classes are differentiated by the 'order' feature in table A.7. The other features ('path' and 'indices') will be illustrated in the next sections. The path in $S C C_{2}$ contains the lexical fillers of the 'AuxP' and 'AuxC' nodes only.

$$
\begin{array}{cc}
\text { SCF } & \text { SCC } \\
\hline S C F_{1}:+ \text { path, +indices, +order } & S C C_{2}:+ \text { path, -indices, -order }
\end{array}
$$

Table A.7: Different types of structures in the SC lexicon.

## A.3.0.1 $S C F_{1}$

As we explained in section A.2.3, certain intermediate nodes - prepositions ('AuxP'), conjunctions ('AuxC') and coordinating ('Coord') or apposing elements ('Apos') - may occur between the verbal headword and the argument nodes. These nodes need to be accounted for in the lexicon, as their presence helps fully describe the distribution of the verbs' syntactic arguments in the corpora. For example, knowing that the verb venio (node venit-12) in sentence (5) (page 60, figure 4.5, page 61) occurs with the object exercitum-11 introduced by the preposition ad-10 is essential to describe which types of prepositions are allowed in the argument structure of the verb venio.

Likewise, for certain purposes (such as searching for coordinated nouns in a corpus) it may be useful to extract the coordinated objects of the verb do in sentence (6) (page 60, figure 4.6, page 62): Senonibus-3 and Gallis-6. For these reasons, we decided to include the intermediate nodes into the most specific type of SCSs: $S C F_{1}$, displayed in table A. 7 .

Special database queries associate each verb-target pair to the tree path formed by the nodes found between the verbal node and the target node (by target node we mean an argument node). The 'path' feature in the first row of table A. 7 refers to this representation. For example, the $S C F_{1}$ for venio in sentence (5) is:

$$
S C F_{1}=(\mathrm{AuxP}) \mathrm{Obj}+\mathrm{V}
$$

The path is represented in brackets before the node it belongs to. When the path contains more than one intermediate node, all these nodes appear in the $S C F_{1}$ structure, in descending order along the tree. If several 'Coord' or 'Apos' nodes occur, they need to be given different indices so that they can be distinguished from each other: the 'indices' feature in table A. 7 refers to this.

The adoption of these indices turns out to be particularly useful to disambiguate SCSs where more coordinated or apposed objects ('Obj_Co' or 'Obj_Ap') refer to different verbal arguments. For instance, let us consider the following sentence, whose tree is represented in figure A.10. In this sentence the pronoun illa refers to the noun ars, previously occurring in the text; reservat is a coordinated predicate of a subordinate adverbial clause.
(4) et illa reservat finem, scilicet
and that:NOM.F.SG reserve:IND.PRS.3SG further:ADV.COMPARATIVE end:ACC.F.SG, namely
usum navis, arti superiori, scilicet gubernatoriae. use:ACC.M.SG ship:GEN.F.SG, art:DAT.F.SG superior:DAT.F.SG, namely of governing:DAT.F.SG. (Th., Sent. P. L., IV, D. 7, Q. 3, Art. 1B, S., 7-4.8-7)
'and that [art] reserves the final end, i. e. the use of the ship, to the superior art, i.e. that of governing.'


Figure A.10: Dependency tree of sentence (4)

In this sentence, two pairs of objects (finem-usum and arti-gubernatoriae) receive the same functional label 'Obj_Ap'. In order to disambiguate these two pairs, they are assigned different indices: the apposing element scilicet, heading the two direct objects finem and usum, is provided with index 1, whereas the second scilicet, heading the two indirect objects arti and gubernatoriae, is provided with index 2.

In addition, $S C F_{1}$ records the full paths from the verbal head V to its argument nodes 'Obj_Ap', as table A. 8 shows. If we provide the SCSs with morphological features, we notice that the two pairs of objects have also different cases: accusative for the first one and dative for the second one (see table A.8).

$$
\begin{aligned}
S C F_{1}: & \\
& S b+V+\left(\text { Apos }_{1}\right) O b j_{\_} A p+\left(\text { Apos }_{1}\right) O b j_{\_} A p+ \\
& +\left(A p o s_{2}\right) O b j_{\_} A p+\left(A p o s_{2}\right) O b j_{\_} A p \\
S C F_{1_{\text {morph }}:} & S b[\text { nom }]+V+\left(\text { Apos }_{1}\right) O b j_{\_} A p[a c c]+\left(\text { Apos }_{1}\right) O b j_{-} A p[a c c]+ \\
& +\left(\text { Apos }_{2}\right) O b j_{\_A p}[d a t]+\left(\text { Apos }_{2}\right) O b j_{\_} A p[d a t]
\end{aligned}
$$

Table A.8: $S C F_{1}$ structure of reservo in sentence (4).

## A.3.0.2 $S C C_{2}$

$S C C_{2}$ structures disregard the linear order of the arguments in the sentences. In cases of coordinated or apposed arguments, $S C C_{2}$ does not record their actual realisation but it creates a syntactic category corresponding to the common functional label of the coordinated or apposed elements. For example, in sentence (4) (page 183), $S C C_{2}$ consists of a subject and an object, even though this abstract object is realised as a pair of coordinated objects in the sentence. This way, $S C C_{3}$ represents the valency of
the verb. The $S C C_{2}$ for the triargumental occurrence of reservo in sentence (4) is 'Obj, $\mathrm{Obj}, \mathrm{Sb}$. It is clear that in this case there is no possible $S C F_{2}$, as the linear order of the sentence cannot be preserved after defining the abstract syntactic categories. ${ }^{14}$

In addition, $S C C_{2}$ makes explicit which preposition or conjunction appears in the path along the tree between the verb and its arguments; in other words, $S C C_{2}$ keeps only those paths that contain 'AuxP' and 'AuxC' nodes. For instance, the $S C C_{2}$ for venio in sentence (5) is '(ad) Obj , Sb '.

## A. 4 Frequency information in the lexicon

In this section we focus on the frequency information contained in the lexicon by showing how to extract it (section A.4.1). Finally, in section A.4.2 we look for a correlation between verb frequencies and SC counts.

## A.4.1 How to extract frequency information

As the SC sublexicons consist of database tables, it is possible to group their rows according to specific criteria. For example, we can calculate the frequency of co-occurrence of each verb lemma with each of the previously introduced SC structures by grouping the rows of the database table containing the lexicon ('subcat_lexicon_db', 'db'='ldt' or 'it') by the (verb, SCS) pairs. As an example, query A.4.1 calculates freq(verb, $\left.S C C_{2_{\text {morph-voice }}}\right)$.

Query A.4.1. SELECT verb, scc2_morph-voice, count(scc2_morph-voice) as
freq_verb_scc2_morph-voice FROM subcat_lexicon_db
group by verb, scc2_morph-voice;

Table A. 9 displays the frequency $\operatorname{freq}\left(d o, S C C_{2_{\text {morph-voice }}}\right)$ in the IT-TB sublexicon and was obtained with query A.4.1 by setting 'verb'='do'. From table A. 9 we can see that two thirds (36) of the 53 occurrences of do are in the active form and associated with at least one of the following arguments: Obj[abl], Obj[dat], and $\mathrm{Sb}[\mathrm{nom}]$, even though they do not always appear together in the same sentence. This reflects the intuition that this verb has three possible argument slots: a nominative subject, an accusative direct object and a dative indirect object. Moreover, this is confirmed by Happ (1976a, page 559), who lists two possible SC frames for do: the first requires two obligatory arguments (nominative and accusative) and the second adds a third dative argument.

Looking at passive usages, the objects in ablative (Obj[abl]) always depend on the preposition $a / a b$ ('by') in our data and are agentive arguments, corresponding to the subjects in the active form. On the other hand, the dative objects (Obj[dat]) correspond to the same arguments occurring with the active form of the verb.

## A.4.2 SC frequency and verb-SC frequency

After presenting the distributions of verbs and SCSs in section 4.7, we present some quantitative data on frequent SCSs (section A.4.2.1) and analyse the number of SCSs


| $S C C_{2_{\text {morph-voice }}}$ | frequency with $d o$ |
| ---: | :--- |
| A_Obj[acc] | 7 |
| A_Obj[acc],Obj[dat] | 3 |
| A_Obj[acc],Obj[dat],Sb[nom] | 5 |
| A_Obj[acc],Sb[nom] | 13 |
| A_Obj[dat],Obj[acc] | 4 |
| A_Obj[dat],Obj[acc],Sb[nom] | 4 |
| P_(a)Obj[abl] | 2 |
| P_(a)Obj[abl],Obj[dat],Sb[nom] | 1 |
| P_(a)Obj[abl],Sb[nom] | 1 |
| P_Obj[dat] | 1 |
| P_Obj[dat],Sb[sbjv] | 2 |
| P_Obj[dat],Sb[inf] | 1 |
| P_Obj[dat],Sb[nom] | 1 |
| P_Pnom[inf],Sb[nom] | 1 |
| P_Sb[nom] | 5 |
| P_V | 2 |

Table A.9: Frequency of the verb $d o$ with its $S C C_{2}$ $\qquad$ structures in the IT-TB lexicon.
per verb (section A.4.2.2).

## A.4.2.1 Frequent SC structures

In this section we show two tables referred to in section 4.7.2. Table A. 10 contains the 5 top-frequency SCCs by author in the lexicon, while table A. 11 shows how the arguments are morphologically realised.

| Caesar | A_V | 35 |
| :---: | :---: | :---: |
|  | A_Obj[acc] | 29 |
|  | A_Obj[acc], $\mathrm{Sb}[\mathrm{nom}$ ] | 22 |
|  | A_Sb[nom] | 21 |
|  | P_Sb[abl] | 15 |
| Cicero | A_V | 216 |
|  | A_Obj[acc] | 205 |
|  | P_V | 141 |
|  | A_Sb[nom] | 96 |
|  | A_Obj[acc],Sb[nom] | 95 |
| Jerome | A_V | 393 |
|  | A_Obj[acc] | 276 |
|  | A_Sb[nom] | 184 |
|  | A_Obj[acc], $\mathrm{Sb}[\mathrm{nom}$ ] | 129 |
|  | P_Sb[nom] | 87 |
| Ovid | A_Obj[acc] | 158 |
|  | A_Obj[acc], $\mathrm{Sb}[\mathrm{nom}$ ] | 141 |
|  | A_V | 135 |
|  | P_V | 108 |
|  | A_Sb[nom] | 99 |
| Petronius | A_V | 452 |
|  | A_Obj[acc] | 442 |
|  | A_Obj[acc], $\mathrm{Sb}[\mathrm{nom}$ ] | 268 |
|  | A_Sb[nom] | 252 |
|  | P_V | 164 |
| Propertius | A_Obj[acc] | 154 |
|  | P_V | 111 |
|  | A_V | 104 |
|  | A_Obj[acc],Sb[nom] | 103 |
|  | A_Sb[nom] | 76 |
| Sallust | A_Obj[acc] | 310 |
|  | A_V | 250 |
|  | A_Obj[acc], $\mathrm{Sb}[\mathrm{nom}$ ] | 238 |
|  | P_V | 184 |
|  | A_Sb[nom] | 177 |
| Vergil | A_Obj[acc] | 94 |
|  | A_V | 84 |
|  | P_V | 62 |
|  | A_Sb[nom] | 58 |
|  | A_Obj[acc],Sb[nom] | 56 |
| Thomas | P_V | 1188 |
|  | A_V | 975 |
|  | A_Pnom[nom],Sb[nom] | 955 |
|  | A_Sb[nom] | 855 |
|  | P_Sb[nom] | 629 |

Table A.10: Five most frequent types of SCC and their frequencies for each author in our SC lexicon.

| treebank | Sb |  | Obj |  | Pnom |  | OComp |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| LDT | case/mood | freq | case/mood | freq | case/mood | freq | case/mood | freq |
|  | nom | 4528 | acc | 4657 | nom | 874 | acc | 71 |
|  | acc | 606 | inf | 1014 | acc | 142 | nom | 5 |
|  | abl | 230 | dat | 907 | gen | 35 | - | 4 |
|  | inf | 143 | abl | 833 | - | 22 | gen | 2 |
|  | ind | 93 | ind | 292 | abl | 19 | abl | 2 |
|  | sbjv | 49 | sbjv | 225 | dat | 16 |  |  |
|  | - | 36 | nom | 91 | inf | 10 |  |  |
|  | gen | 24 | imp | 76 | ind | 7 |  |  |
|  | voc | 2 | gen | 57 | sbjv | 3 |  |  |
|  | dat | 2 | - | 57 | voc | 1 |  |  |
|  | imp | 1 | voc | 3 |  |  |  |  |
|  |  |  | loc | 2 |  |  |  |  |
|  |  |  | supine | 1 |  |  |  |  |
| IT-TB | nom | 4536 | acc | 1912 | nom | 1709 | acc | 35 |
|  | sbjv | 240 | abl | 553 | inf | 99 | abl | 3 |
|  | acc | 160 | inf | 517 | gen | 81 | ind | 1 |
|  | ind | 137 | dat | 450 | acc | 60 |  |  |
|  | inf | 128 | ind | 132 | ind | 34 |  |  |
|  | abl | 73 | sbjv | 37 | - | 4 |  |  |
|  | - | 5 | gen | 27 | sbjv | 1 |  |  |
|  | imp | 2 | nom | 4 | abl | 1 |  |  |
|  | adv | 1 | adv | 3 | dat | 1 |  |  |
|  | dat | 1 | - | 2 |  |  |  |  |
|  |  |  | imp | 1 |  |  |  |  |

Table A.11: Morphological realisations of verbal arguments in the lexicon. Column 'Sb': the rows headed 'acc' and 'abl' refer to subjects in accusative + infinitive constructions and in ablative absolutes, respectively; 'ind' and 'sbjv' refer to subject clauses whose subjects are expressed as verbs in the indicative and subjunctive mood, respectively. For the other cases, see the text.

## A.4.2.2 An analysis on the number of structures per verb

Table A. 12 contains the counts of the different SC structures for the ten most frequent verbs. We now want to explore the possibility of a correlation between the verb

| treebank | verb | frequency | $S C F_{1_{\text {voice }}}$ | $S C C_{2_{\text {voice }}}$ |
| :--- | :---: | :---: | ---: | ---: |
| LDT | sum | 1535 | 108 | 36 |
|  | habeo | 308 | 63 | 32 |
|  | dico | 268 | 75 | 29 |
|  | facio | 261 | 86 | 35 |
|  | video | 247 | 58 | 23 |
|  | possum | 179 | 29 | 7 |
|  | do | 154 | 52 | 24 |
|  | venio | 128 | 20 | 15 |
|  | inquam | 94 | 19 | 7 |
| IT-TB | audio | 68 | 20 | 11 |
|  | sum | 2475 | 79 | 20 |
|  | dico | 795 | 80 | 33 |
|  | habeo | 372 | 48 | 18 |
|  | possum | 318 | 24 | 7 |
|  | facio | 205 | 28 | 15 |
|  | video | 152 | 30 | 19 |
|  | oportet | 132 | 11 | 7 |
|  | recipio | 132 | 25 | 15 |
|  | pono | 129 | 47 | 26 |
|  | debeo | 101 | 17 | 5 |

Table A.12: Ten most frequent verbs in the LDT and in the IT-TB, and counts of their SCSs in the lexicon.
frequency and the counts of the different SCSs per verb.
Figures A. 11 and A. 12 show the raw verb frequencies plotted against the number of SCSs per verb. Based on these plots, there seems to be a positive non-linear correlation between the two variables. ${ }^{15}$ To test for a correlation between the verb frequencies and the counts of the different SCSs per verb, we used the Spearman rank correlation test computed in R (R Development Core Team, 2008). This test is an ordinal version of the Pearson correlation test: it converts the frequencies to ranks and, unlike the latter, is very robust against outliers and non-linear correlations. The test results are provided in table A.13. All the $p$-values, which represents the likelihood of the rank pairs arising by chance, are significant at the 0.01 level. ${ }^{16}$ The $r_{s}$ coefficients indicate a strong positive correlation for all structures. In particular, the structure that takes the sentence order into account $\left(S C F_{1}\right)$ displays a slightly higher correlation $\left(0.94<r_{s}<0.96\right)$ than the one that does not $\left(S C C_{2}, 0.92<r_{s}<0.93\right)$, and the correlation coefficients are higher for the LDT than for the IT-TB.

Since the corpora are finite, our results are indeed partially influenced by lowfrequency verbs, which occur with very few frames. For example, if a verb is only seen once in the corpora, the number of its frames is 1 ; however, this does not mean

[^68]

Figure A.11: Verb frequencies and number of SCSs in the LDT lexicon. See table A. 12 for the ten most frequent verbs.


Figure A.12: Verb frequencies and number of SCSs in the IT-TB lexicon. See table A. 12 for the ten most frequent verbs.

| treebank | SC | full dataset |  |  | freq. $>26$ |  |  |
| :--- | :--- | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | $r_{s}$ | $S$ | $p$-value | $r_{s}$ | $S$ | $p$-value |
| LDT | $S C F_{1}$ | 0.96 | 31247864 | $<0.01$ | 8461 | 0.68 | $<0.01$ |
|  | $S C C_{2}$ | 0.93 | 50770160 | $<0.01$ | 11719 | 0.55 | $<0.01$ |
| IT-TB | $S C F_{1}$ | 0.94 | 1615016 | $<0.01$ | 11098 | 0.55 | $<0.01$ |
|  | $S C C_{2}$ | 0.92 | 2284285 | $<0.01$ | 10278 | 0.59 | $<0.01$ |

Table A.13: Correlation coefficient $r_{s}$, test statistic $S$ (the sum of rank differences) and $p$-value for each correlation test.
that the verb could not appear with other frames in a larger corpus. This could partially explain why the correlation between the verb frequency and the frame count is so high.

To see how hapaxes and low-frequency verbs affected our results, we performed the same analysis on the 54 (LDT) and 53 (IT-TB) most frequent verbs only (with frequency> 26) and found lower, but still medium-sized correlations, whose results are reported in the last three columns of table A.13. These results disagree with the ones presented in Hinrichs and Telljohann (2009), where the authors - without making use of correlation tests - find a weak correlation between the two variables for a much larger corpus ( 36,000 sentences) and for the most frequent verbs only.

## Appendix B

## Statistical background to chapter 5

This appendix gives various definitions, mathematical details and data that are useful for a more in-depth understanding of chapter 5 .

In sections B. 1 and B. 2 we will touch upon basic definitions in Bayesian statistics and Information Theory. Sections B. 3 and B. 6 give a more detailed description of the previous systems for SP acquisition, while section B. 4 introduces multidimensional spaces and B. 5 illustrates the main ideas behind cluster analysis. Section B. 7 displays the complete LWN synsets and the semantic properties we refer to in chapter 5. Finally, section B. 8 presents the formulae we designed for our probabilistic model for SP acquisition described in chapter 5 .

## B. 1 Bayesian statistics

In this section we will briefly summarise the main concepts in Bayesian inference to the extent that they were employed in our SP model described in chapter 5. For a fuller illustration of these concepts in a data-analysis perspective we refer to Gelman et al. (2004a).

## B.1.1 Prior probability and posterior probability

As other types of statistical inference, Bayesian inference involves making inferences from a sample to the whole population. In Bayesian inference the data of the study are used to update the probability of a hypothesis. This is done by comparing the probability of a hypothesis $H$ before some evidence $D$ has been collected for it (prior probability) with the probability of the hypothesis after evidence has been observed (posterior probability). Prior and posterior probabilities can be formalised as $P(H)$ and $P(H \mid D)$, respectively.

Conditional probability The probability $P(H \mid D)$ is the called conditional probability of $H$ given $D$ and represents the probability of $H$ resulting given that $D$ has been observed. $P(H \mid D)$ is different from $P(H)$ if $H$ and $D$ are not independent events, that is, if the probability of $H$ after observing $D$ is different than before observing $D$.

The conditional probability $P(H \mid D)$ is defined as:

$$
P(H \mid D)=\frac{P(H \cap D)}{P(D)}
$$

where the joint probability of $H$ and $D$, i. e. $P(H \cap D)$, is the probability of $H$ and $D$ occurring together.

## B.1.2 Bayes' Theorem

Bayesian inference uses Bayes' theorem to update the posterior probability from the prior probability:

$$
\begin{equation*}
P(H \mid D)=\frac{P(D \mid H) P(H)}{P(D)} \tag{B.1}
\end{equation*}
$$

As can be seen from B. $1 P(H)$ is multiplied by the factor $\frac{P(D \mid H)}{P(D)}$ to obtain $P(H \mid D)$ : this factor is greater than 1 if it is likely that the data $D$ would be observed when $H$ is true, but unlikely that $D$ would have been the outcome of the observation; otherwise, $P(H)$ is multiplied by a number lower than one, resulting in $P(H \mid D)<P(H)$.

## B. 2 Entropy

In this section we will present an important information-theoretic measure widely used in Computational Linguistics: entropy.

If $X$ is a random variable, i. e. a mapping from a sample space composed of events to real numbers, the entropy of $X$ can be defined as the amount of uncertainty associated with $X$. For example, a fair coin flip corresponds to a probability of 0.5 for head and the same probability for tail; this is associated with the concept of maximum entropy, since we cannot predict the outcome of a specific toss. With a non-fair coin that always comes up head, the entropy is 0 because there is no uncertainty.

Entropy can also be considered in relation with message coding as the amount of bits of information contained in a message in bits (Shannon's entropy). For example, each time a fair coin is tossed, it delivers one bit (maximum) of information; a non-fair coin delivers less than a bit, since the outcome is more predictable.

If we consider a discrete random variable $X$ taking the $n$ values $x_{1}, \ldots x_{n}$ and the probability mass function $p$ of $X,{ }^{1}$ then the formula for the entropy of $X$ is:

$$
H(X)=-\sum_{i=1}^{n} p\left(x_{i}\right) \log p\left(x_{i}\right)
$$

The Principle of Maximum Entropy states that, subject to known constraints called testable information, the probability distribution with the highest entropy is the distribution which best represents the current state of knowledge. If there is no testable information, the maximum entropy probability distribution is the uniform distribution, assigning $p_{i}=\frac{1}{n}$ to each event $i \in\{1, \ldots, n\}$.

[^69]
## B. 3 Knowledge-based approaches in selectional preference acquisition

In this section we will give some details of previous knowledge-based systems for SP acquisition introduced in section 5.2.1 and briefly described in section 5.2.1.3.

## B.3.1 Resnik (1993)

Resnik's system populates the WN hierarchy by evenly distributing the frequency of a noun among all the classes it (directly and indirectly) belongs to. Therefore, his approach does not distinguish between hyponymy and polysemy.

Let $v$ be a verb and $r$ and the argument position we are interested in. Resnik distributes the number of observations of a noun uniformly across all the classes containing it. The frequency of a class (synset) $c$ is hence:

$$
f r(c)=\sum_{n} \frac{f r(n)}{|\operatorname{classes}(n)|},
$$

where the sum is over all nouns $n$ belonging to $c$ or to synsets lying under $c$ in the hierarchy, and $|\operatorname{classes}(n)|$ is the number of classes into which $n$ belongs.

| noun | frequency |
| :---: | :---: |
| bee | 3 |
| bug | 0 |
| bird | 4 |
| crow | 2 |
| eagle | 4 |
| insect | 0 |
| swallow | 0 |

Table B.1: Example corpus frequencies for the subjects of fly.


Figure B.1: Example portion of a WN-like hierarchy for some possible subjects of fly listed in table B.1.

In the example represented in table B. 1 and relative to figure B.1, the frequency of
the class animal $=\{\text { animal }\}^{2}$ is:

$$
\begin{array}{r}
f r(\text { ANIMAL })=\frac{f r(\text { swallow })}{\mid \text { classes }(\text { swallow }) \mid}+\frac{f r(\text { crow })}{\mid \text { classes }(\text { crow }) \mid}+\frac{f r(\text { eagle })}{\mid \text { classes }(\text { eagle }) \mid}+ \\
+\frac{f r(\text { bird })}{\mid \text { classes }(\text { bird }) \mid}+\frac{f r(\text { bug })}{\mid \text { classes }(\text { bug }) \mid}+ \\
+\frac{f r(\text { bee })}{\mid \text { classes }(\text { bee }) \mid}+\frac{f r(\text { insect })}{\mid \text { classes }(\text { insect }) \mid}= \\
\\
=0+\frac{2}{3}+\frac{2}{3}+\frac{4}{3}+0+\frac{2}{3}+0=\frac{10}{3}
\end{array}
$$

Similarly, $f r(\operatorname{BIRD})=0+\frac{2}{3}+\frac{2}{3}+\frac{4}{3}=\frac{8}{3}$ and $f r($ INSECT $)=0+\frac{2}{3}+0=\frac{2}{3}$.
Finally, for the leaf node classes: $f r($ swallow $)=f r($ bug $)=f r($ insect $)=0 ; f r($ crow $)=$ $\frac{2}{3}, f r(e a g l e)=\frac{2}{3}, f r($ bee $)=\frac{2}{3}$ and $f r($ bird $)=\frac{4}{3}$.

To estimate the probability of a class $c$ based on its frequency distribution, Resnik uses maximum likelihood (MLE) estimates:

$$
P(c)=\frac{f r(c)}{N}
$$

where $N$ is the total size of the corpus.
In the previous example, $P($ ANIMAL $)=\frac{10}{30}=\frac{1}{3}, P($ Bird $)=\frac{4}{15}, P(\operatorname{INSECT})=\frac{1}{15}$, $P($ swallow $)=0, P($ crow $)=\frac{1}{15}, P($ eagle $)=\frac{1}{15}, P($ bird $)=\frac{2}{15}, P($ bug $)=0, P($ bee $)=\frac{1}{15}$ and $P($ insect $)=0$. We can easily verify that $\sum_{c \text { class }} P(c)=1$.

Resnik proposes to measure the semantic fit of a WN class $c$ as an argument to a verb $v$ in an argument relation $r$ through the concept of selectional association $A(v, c)$, defined as follows: ${ }^{3}$

$$
A_{r}(v, c)=\frac{P(c \mid v) \log \frac{P(c \mid v)}{P(c)}}{S_{r}(v)}
$$

The normalising factor $S_{r}(v)$ is called selectional preference strength of $v$ and is defined as:

$$
S_{r}(v)=\sum_{c} P(c \mid v) \log \frac{P(c \mid v)}{P(c)}
$$

As can be noted, the selectional association scales the mutual information (cf. section B.2) by the conditional probability $P(c \mid v)$. The selectional preference strength coincides with the relative entropy (Kullback and Leibler, 1951) between the prior distribution $P(c)$ and the posterior distribution $P(c \mid v)$ (cf. definitions in section B.1). It measures the influence of $v$ on its arguments, or equivalently, it is the cost of ignoring the verb when estimating the information carried by the predicate about its arguments.

Finally, given a predicate $v$, Resnik's method calculates the association scores of all candidate noun classes for $v$ and then chooses the noun class that has the maximum association score: therefore, the selectional association between $v$ and its argument $n$ is

$$
A_{r}(v, n)=\max _{c \in \operatorname{classes}(n)} A(v, c)
$$

[^70]
## B.3.2 Ribas (1995)

Ribas (1995) proposes to compute the frequency of a class $c$ in the WN population step by summing the weighted frequencies of all nouns contained in $c$ or in classes under $c$ :

$$
f r(c)=\sum_{n \text { noun } \in \text { cor under } c} f r(n) \cdot \text { weight }(n, c),
$$

where the weight is defined as the number of direct classes for $n$ that are under $c$ including $c$, divided by the total number of direct classes for $n$.

In the example above (see table B.1), fr(swallow) $=f r(b u g)=0, f r(c r o w)=$ $2 \cdot \frac{1}{1}=2, f r($ eagle $)=2 \cdot \frac{1}{1}=2, f r($ bird $)=4 \cdot \frac{1}{2}=2, f r(b u g)=0, f r($ bee $)=2 \cdot \frac{1}{1}=2$, $f r($ insect $)=0, f r(\operatorname{BIRD})=0+2 \cdot \frac{1}{1}+2 \cdot \frac{1}{1}+4 \cdot \frac{2}{2}=8, f r(\operatorname{INSECT})=0+3 \cdot \frac{1}{1}+0=3$, and $\operatorname{fr}($ ANIMAL $)=0+2 \cdot \frac{1}{1}+2 \cdot \frac{1}{1}+4 \cdot \frac{2}{2}+0+2 \cdot \frac{1}{1}+0=10 .{ }^{4}$

The corresponding probabilities are computed as relative frequencies through MLE: $P($ swallow $)=0, P($ crow $)=\frac{2}{10}=\frac{1}{5}, P($ eagle $)=\frac{1}{5}, P($ bird $)=\frac{1}{5}, P($ bug $)=0$, $P($ bee $)=\frac{1}{5}, P($ insect $)=0, P(\operatorname{BIRD})=\frac{4}{5}, P(\operatorname{INSECT})=\frac{3}{10}, P($ ANIMAL $)=\frac{10}{10}=1$.

Note that, unlike in Resnik's model, the top node has probability 1 and that the sum of the probabilities of all senses of a noun $n$ - rather that the sum of all senses and hypernym classes, as in Resnik's model - is 1 :

$$
\sum_{\text {s sense of } n} P(s \mid n)=1
$$

For example, $\sum_{s \text { sense of bird }} P(s \mid$ bird $)=P($ bird $)+P(\operatorname{BIRD})=\frac{2}{10}+\frac{8}{10}=1$.
In addition, Ribas's investigates the use of log-likelihood ratio (Dunning 1993) in calculating the association score, as well as the limited impact of thresholding. Finally, as far as ambiguity is concerned, Ribas proves that using WN senses - in particular the most frequent one or the sense that is most associated with the complements of the ambiguous noun - leads to better results that assigning basic senses to nouns.

## B.3.3 Li and Abe (1998)

Li and Abe $(1995,1998)$ propose a generalization method based on the Minimum Description Length (MDL) principle, a principle of data compression from Information Theory. Their data are argument pairs extracted from the Wall Street Journal corpus.

Li and Abe's method works on thesaurus trees, i.e. trees in which each leaf node stands for one and only one noun and each internal node stands for the class of nouns below it; an example is represented in figure B.1. The noun taxonomy of WN is structured as directed acyclic graphs (DAGs), where the nodes are connected to each other by the IS-A relation, and stand for a word sense but often not for one noun only. To adapt WN to their model, Li and Abe obtain a tree structure from WN through an algorithm that transforms DAGs into tree structures; this algorithm prunes the hierarchy at classes where a direct class member (one of the nouns in the synset) has occurred in the data. They then equally divide the observed frequency of a noun between all the nodes containing that noun, in order to heuristically deal with word sense ambiguity.

[^71]After this operation, SPs are represented as cuts, i. e. partitions which disjointly cover all the leaf nodes across the hierarchy. The generalization problem is then reduced to that of estimating a tree cut model (TCM) of the thesaurus tree. For example, ${ }^{5}$ let us suppose that table B. 1 contains the frequencies of the fillers for the subject slot of the verb fly. The thesaurus tree is represented in figure B.1; in this tree there are five possible cuts: [ANIMAL] (the most general one), [BIRD, insECT], [BIRD, bug, bee, insect], [swallow, crow, eagle,bird, inSECT], and [swallow, crow, eagle, bird, bug, bee, insect] (the most specific one).

The best model is choosen in the sense of MDL: in other words it is the one which minimises the Total Description Length $t d l$, that is the sum of:

- the Model Description Length $(m d l)$ : number of bits for the encoding of the model itself;
- the Data Description Length $(d d l)$ : number of bits for the encoding of the given data observed through the model.
Because $m d l$ is related to the model's complexity and $d d l$ indicates how well the model fits the data, MDL identifies the optimal analysis of the corpus as the one providing the most compact and still accurate representation of the data.

For the example above, as can be seen in table B.2, a model nearer the root of the thesaurus tree is simpler in terms of the number of parameters, but has a poorer fit to the data; on the other hand, a model nearer the leaves of the thesaurus tree is more complex, while having a better fit to the data. Therefore, the algorithm by Li and Abe,

| model | $d d l$ | $m d l$ | $t d l$ |
| :--- | :---: | :---: | :---: |
| [ANIMAL] | 28.07 | 28.07 | 56.14 |
| [BIRD, INSECT] | 26.39 | 29.71 | 30.43 |
| [BIRD, bug, bee, insect] | 23.22 | 33.18 | 56.4 |
| [swallow, crow, eagle, bird, INSECT] | 22.39 | 26.67 | 49.06 |
| [swallow, crow, eagle, bird, bug, bee, insect] | 19.22 | 39.16 | 58.38 |

Table B.2: Description lengths of the five TCMs for the thesaurus in figure B. 1 ( Li and Abe, 1998).
in this case, returns the second model ([BIRD, INSECT] ) as the best one.

## B.3.4 McCarthy (2001)

McCarthy (2001) adopts an approach similar to that proposed by Li and Abe (1995, 1998), but she enriches the preference acquisition with a WSD system.

McCarthy compares the results obtained in two different experiments: the first one uses Li and Abe's original system, while in the second one the nouns in the target slots were previously disambiguated: the WSD system is that proposed by Wilks and Stevenson (1998) in which the most frequent sense is extracted regardless of context. McCarthy shows that even this very simple kind of disambiguation improves the SP acquisition process, especially when dealing with sparse data. She also obtains promising results trying an iterative approach, where an initial WSD is applied to acquire SPs, which are again used to disambiguate the heads and help the following SP acquisition step.

[^72]
## B.3.5 Clark and Weir (1999)

Clark and Weir (1999) show that the usual technique which splits the frequency counts of a noun appearing in the data evenly among its alternative senses can cause inaccurate estimates when dealing with ambiguous data. For this reason, they describe a re-estimation process where the cut moves up in the WN noun hierarchy.

Given a verb $v$ and a relation $r$, the starting model distributes the counts for a noun uniformly among all its senses. They assume that lower-level sibling concepts tend to accumulate at the best hypernym of a class $c$, that is the hypernym that best represents the set of related senses of arguments for a verb $v$ in relation $r$. The decision about coalescing two nodes into the parent node is made by applying the $\chi^{2}$ test on the null hypothesis that the probability of a hypernym is independent of whether it is an element of a particular child class (see section C. 2 for an illustration of the $\chi^{2}$ test).

After defining the tree cuts, they calculate the frequency of a class $c$ by weighting the frequencies of the nouns contained in it by association norms, defined this way:

$$
A(c, v, r)=\frac{P(c \mid v, r)}{P(c \mid r)} .
$$

The experiments performed by the authors prove that this re-estimation procedure makes the estimates more accurate, especially for verbs which strongly select for their objects.

## B.3.6 Abney and Light (1999)

Abney and Light (1999) introduce a model for SP induction based on Hidden Markov Models (HMMs), a sort of "stochastic version of a non-deterministic finite state machines" (Light and Greiff, 2002, p. 277).

Contrary to the usual uniform distribution for ambiguous nouns, in this model each sense of a word has its own probability distribution, because each pair $(v, r)$ is associated to a HMM, whose shape is identical to that of WN. States and transitions of a HMM correspond to nodes (semantic classes) and arcs in the hierarchy (IS-A relations), and a run of one of these HMMs begins at the root of the hierarchy. Transitions from a class to its child class are done iteratively using the HMM's transition probabilities, until a terminal node is reached and a word to express the given sense is generated.

Abney and Light estimate parameters for each HMM by means of a forward-backward algorithm, from sample observation sequences consisting of nouns that occurred with certain verbs. Unfortunately, despite some smoothing and balancing tecniques, one of the disadvantages of Abney and Light's method is related to the parameter estimation and this causes negative results in disambiguating correctly the senses of input words.

## B.3.7 Ciaramita and Johnson (2000)

Ciaramita and Johnson (2000) use Bayesian Belief Networks (BBNs) for inferring the probability that a verb selects for a particular class in the network.

A BBN (Pearl, 1988) is identified by DAGs, defined by sets of variables (nodes, associated to a finite number of mutually exclusive states) and sets of directed edges which connect the variables. For each verb, Ciaramita and Johnson's method creates a BBN whose architecture looks like WN, whose parameters are estimated from the
occurrences of verb-object pairs in the corpus and whose variables are all boolean. ${ }^{6}$ They then combine the prior knowledge encoded in the network with the observed data and Bayesian inference.

In Ciaramita and Johnson's networks, the prior probability distributions follow these intuitive principles.

- it is unlikely that a verb a priori selects for any particular class;
- if a verb selects for a class, then it is likely that it also selects for its hypernyms;
- if a verb selects for any of the possible senses of a word, then it is likely that it selects for that word;
- if a verb does not select for a class, then it is unlikely that it selects for the children of that class.

The authors test their method on a WSD task, getting very good results when compared to the state-of-the-art systems.

## B.3.8 Agirre and Martinez (2001)

The input data in the system by Agirre and Martinez (2001) are (noun sense, relation, verb sense) triples ( $c n, r, c v$ ) from the SemCor corpus (Miller et al., 1993), a portion of the Brown corpus (Francis, 1979) consisting of 250,000 words manually tagged with WN senses.

The probabilities for the noun classes $c n_{i}$ that are in relation $r$ (subject or object) to $v$ are calculated this way:

$$
\begin{gathered}
P\left(c n_{i} \mid r, v\right)=\max _{c v_{j} \text { senseof } v} \sum_{c n \supseteq c n_{i}} \sum_{c v \supseteq c v_{j}} P\left(c n_{i} \mid c n\right) \cdot P\left(c v_{j} \mid c v\right) \cdot P(c n \mid r, c v)= \\
=\max _{c v_{j} \text { senseof } v} \sum_{c n \supseteq c n_{i}} \sum_{c v \supseteq c v_{j}} \frac{\hat{f r}\left(c n_{i}, c n\right)}{f r(c n)} \cdot \frac{\hat{f r}\left(c v_{j}, c v\right)}{\hat{f r}(c v)} \cdot \frac{\hat{f r}(c n r, c v)}{f r(r, c v)}
\end{gathered}
$$

where $\hat{f r}$ is the estimated frequency and

$$
\hat{f r}(c n)=\sum_{c n_{i} \subseteq c n} \frac{f r\left(c n_{i}\right)}{\left|\operatorname{classes}\left(c n_{i}\right)\right|}
$$

For example, the probability of chicken $_{1}$ (i.e. the first sense of chicken) as an object for eat (i.e. the verb class \{ingest, take in...\}) depends on the probabilities that all concepts above chicken $1_{1}$ are objects of all concepts above the possible senses of eat.

Agirre and Martinez test their system on a WSD task and prove that a class-to-class approach gets the best recall when compared with word-to-word and word-to-class approaches, which have better precision but low coverage.

[^73]| verb token | verb lemma | author |
| :---: | :---: | :---: |
| $v_{1}$ | dico | Plautus |
| $v_{2}$ | $d o$ | Cicero |
| $v_{3}$ | dico | Caesar |
| $v_{4}$ | $d o$ | Caesar |

Table B.3: Constructed example of dataset recording the author and the verb lemma for each verbal occurrence observed.

| author/verb | dico | do |
| :--- | :---: | :---: |
| Caesar | 1 | 1 |
| Cicero | 0 | 1 |
| Plautus | 1 | 0 |

Table B.4: Frequency table of the dataset represented in table B.3.

## B. 4 Vector spaces

In this section we briefly introduce vectors in multidimensional spaces, since this is a necessary notion for understanding sections B. 5 and C.4. We will first show the data format used in data analysis (section B.4.1) and then give a geometrical definition of vectors (section B.4.2).

## B.4.1 Data format

In data analysis we usually deal with data tables: rows correspond to observations and column to variables. For example, we can collect the occurrences of verbs in a corpus (observations) and, for each of them, record the author whose work the verb occurs in, as well as the verb's lemma (variables 'author' and 'verb').

Table B. 3 shows some constructed data for this example. The cells of the table correspond to the values of each variable relative to each observation; for example, the cell in the first row and third column (dico) records the author of the verbal token $v_{1}$, i. e. Plautus.

Tables like B. 3 can be turned into frequency tables - also called contingency tables. In the case of two variables (in the previous example 'author' and 'verb'), we can assign rows to the values for the first variable ('author') and columns to the values of the second variable ('verb'). We then categorize the observations by these two variables and count the frequency of each variable pair. In the previous example, the values for the first variable 'author' are: Caesar, Cicero and Plautus, while the values for the second variable 'verb' are dico and do. Table B. 4 contains these frequencies.

## B.4.2 Vectors

The rows (and the columns) in frequency tables like B. 4 have a geometrical interpretation as vectors. A vector $v$ in a $n$-dimensional space is identified by an ordered set of $n$ numbers called coordinates of $v: v=\left(v_{1}, \ldots, v_{n}\right)$.

For example, a two-dimensional space can be thought of as the well-known Cartesian plane, defined by two perpendicular lines called $x$ axis and $y$ axis and meeting in the

|  | $c_{1}$ | $c_{2}$ | $\ldots$ | $c_{n}$ |
| :---: | :---: | :---: | :---: | :---: |
| $r_{1}$ | $f r\left(r_{1}, c_{1}\right)$ | $\ldots$ | $f r\left(r_{1}, c_{n}\right)$ |  |
| $\vdots$ | $\ldots$ | $\ldots$ | $\ldots$ |  |
| $r_{m}$ | $f r\left(r_{m}, c_{1}\right)$ | $\ldots$ | $f r\left(r_{m}, c_{n}\right)$ |  |

Table B.5: $m \times n$ table, with $m$ rows (observations) and $n$ columns (variables).
origin 0 . In such a space a vector is associated with a pair of coordinates, called abscissa and ordinate. In the previous example, the points represented in table B. 4 correspond to the first, second and third row and have two coordinates, associated with the two verbs dico and $d o$; these points are plotted into a Cartesian plane in figure B.2: 'first' corresponds to the first row of table B.4, 'second' to the second one, and so on.


Figure B.2: 2-dimensional Cartesian space displaying the points corresponding to the rows of table B.4.

We can generalize this case to a higher-dimensional space with $n$ dimensions, defined by $n$ orthogonal axes. This allows us to work on cases when more than two values for the second variables are observed. For example, we could think of a study where more than two authors are considered. The general form of a table containing $m$ vectors $r_{1}, \ldots, r_{m}$ in a $n$-dimensional space is represented in table B.5: each coefficient in the position $(i, j)$ contains the frequency of observation $r_{i}$ relative to variable $c_{j}$.

## B. 5 Clustering

The term clustering in data analysis denotes a group of statistical techniques sharing the aim of assigning an input set of observations into subsets called clusters; this is done in an unsupervised way, since no human intervention is required. Additionally,
the clusters constitute the output of the analysis and are not known in advance. ${ }^{7}$ In other words, clustering is a way to investigate a dataset from the point of view of the possible coherent subsets it is composed of.

The way this grouping is performed is by representing the observations in the form of row vectors in a table (see section B.4.2) and considering their similarities with respect to the table columns, which are the values of the variables according to which we cluster.

## B.5.1 Overview

In this section we illustrate the main clustering methods used in distributional semantics, for which we give a toy example. For a more comprehensive discussion on uses of clustering in linguistics, we refer to Baayen (2008, pp. 138-148) and Bilisoly (2008, pp. 219-235).

Distance measures The assignment of clusters depends on the way similarity between observations is defined, so that similar observations are grouped together into the same cluster. The common way to define similarity between observations is through some kind of distance measure. ${ }^{8}$

The most intuitive example of distance is the Euclidean distance, whose version in two dimensions is:

$$
\begin{equation*}
d\left(v_{1}, v_{2}\right)=\sqrt{\left(x_{1}-x_{2}\right)^{2}+\left(y_{1}-y_{2}\right)^{2}} \tag{B.2}
\end{equation*}
$$

where $v_{1}=\left(x_{1}, y_{1}\right)$ and $v_{2}=\left(x_{2}, y_{2}\right)$ are two points in a two-dimentional space. ${ }^{9}$
A distance widely used in computational semantics is the cosine distance:

$$
\begin{equation*}
d\left(v_{1}, v_{2}\right)=1-\cos \left(v_{1}, v_{2}\right) \tag{B.3}
\end{equation*}
$$

where $\cos \left(v_{1}, v_{2}\right)$ is the cosine between $v_{1}$ and $v_{2}$ treated as vectors; in two dimensions the cosine is ${ }^{10}$

$$
\begin{equation*}
\frac{x_{1} x_{2}+y_{1} y_{2}}{\sqrt{\left(x_{1}^{2}+y_{1}^{2}\right) \cdot\left(x_{2}^{2}+y_{2}^{2}\right)}} \tag{B.4}
\end{equation*}
$$

The cosine of the angle $\alpha$ between $v_{1}$ and $v_{2}$ can be defined by referring to the unit circle, i. e. a circle with a radius of one: if we consider the line through the origin making the angle $\alpha$ with the positive half of the $x$-axis and we focus on the intersection $P$ between this line and the unit circle, then the cosine of $\alpha$ is the abscissa of $P$. For example, figure B. 3 represents the vectors $v_{1}=(2,1)$ (blue) and $v_{2}=(2,0)$ (red), the angle $\alpha$ and the point $P$.

Since the cosine function takes values between $-1\left(\right.$ for $\left.\alpha=0^{\circ}\right)$ and $+1\left(\alpha=180^{\circ}\right)$, being 0 for $\alpha=90^{\circ}$, the definition in B. 3 implies that this distance is 0 when $v_{1}=v_{2}$ and is 1 when $v_{1}$ and $v_{2}$ are orthogonal. Unlike the Euclidean distance, the cosine distance is affected only by the angle between the vectors and not by their length, i. e. it is norm-invariant.

[^74]

Figure B.3: Unit circle, the two vectors (4,0) and (2,1), the angle $\alpha$ between them and its cosine.

| noun | with fly | with buzz |
| ---: | :---: | :---: |
| bee | 3 | 5 |
| bug | 2 | 2 |
| bird | 4 | 1 |
| crow | 2 | 0 |
| eagle | 4 | 0 |
| insect | 5 | 10 |
| swallow | 2 | 1 |

Table B.6: Constructed frequencies of the subjects of $f l y$ and $b u z z$.

Frequencies from a toy example on the subjects of $f l y$ and buzz are shown in table B. 6 and plotted in figure 5.2 on page 89, whereas table B. 7 reports the Euclidean and cosine distances between the pairs crow-eagle, bird-eagle and bug-swallow.

| noun pair | Euclidean distance | cosine distance |
| ---: | :---: | :---: |
| crow-eagle | 2 | 0 |
| bird-eagle | 1 | 0.03 |
| bug-swallow | 1 | 0.05 |

Table B.7: Euclidean and cosine distance measures between three pairs of nouns from table B. 6.

As can be observed from table B.7, according to the Euclidean distance, bird and eagle are closer together than crow and eagle, whereas the opposite holds for the cosine distance. This has to do with the fact that the cosine distance is norm-invariant: in fact, the cosine distance between two nouns identified by their absolute frequencies equals the cosine distance between their relative frequencies. For an illustration of this claim on the previous example, see table 5.3 on page 90 and figure 5.3 on page 89 , which show the same nouns plotted as different points depending on whether they are identified in the space by their absolute or relative frequencies.

There is a number of other distance measures used in computational linguistics.

Here we will only mention two more: Jaccard coefficient and Dice coefficient, which work on pairs of sets $A$ and $B$ by comparing their sizes:

$$
\begin{align*}
& \operatorname{Jaccard}(A, B)=\frac{|A \cap B|}{|A \cup B|}  \tag{B.5}\\
& \operatorname{Dice}(A, B)=\frac{2|A \cap B|}{|A|+|B|} . \tag{B.6}
\end{align*}
$$

## B.5.2 Hard agglomerative hierarchical clustering

Clustering methods can be classified in several ways. For example, we can distinguish hard clustering from soft (or fuzzy) clustering, depending on whether overlaps between clusters are allowed or not. In addition, partitional clustering differs from hierarchical clustering in the algorithm used: the former determines the number of clusters at the beginning, while the latter proceeds by merging groups together (agglomerative or bottom-up clustering) or splitting them (divisive or top-down clustering). Here, we will focus on hard agglomerative hierarchical clustering and illustrate it on the example given in table B.6, while omitting the description of more complex types of clustering.

In agglomerative hierarchical clusterings, the first step assigns one cluster to each input observation (object). Then, at each step the closest clusters are merged together until one cluster is obtained at the end. ${ }^{11}$ This type of clustering produces a tree representation of the clusters detected in the input, usually called dendrogram: the special layout of the output refers to the multilevel hierarchy highlighted by the cluster analysis.

Two clusters are merged by calculating the distance between them and this is achieved in several different ways: for example, the distance can be obtained as the minimum distance between the elements of each cluster (single linkage algorithms), their maximum distance (complete linkage), their average distance (average linkage), and so on.

By way of example, let us consider again the data from table B.6. We will proceed to treat these data in the following way:

1 calculate the distances between all pairs of objects (nouns);
2 group the objects into a cluster tree;
3 determine where to divide the tree into clusters.

Calculating the distances First, we calculate the distances between all pairs of nouns: in this case we decided to employ the cosine distance. Table B. 7 displays the cosine distances between the points in table B.6.

Grouping the objects The first step of the clustering defines one cluster for each object: cluster $_{1}=\{$ bee $\}, \ldots$, cluster $_{7}=\{$ swallow $\}$.

Subsequently, we need to group the observations: to do this, a linkage function must be chosen. In our case we chose the simplest one, the single function, which calculates the shortest distance between the clusters. According to table B. 8 the closest pair is crow-eagle, which will then constitute the first cluster.

[^75]| noun pair | cosine distance |
| ---: | :---: |
| bee-bug | 0.03 |
| bee-bird | 0.29 |
| bee-crow | 0.49 |
| bee-eagle | 0.49 |
| bee-insect | 0.003 |
| bee-swallow | 0.16 |
| bug-bird | 0.14 |
| bug-crow | 0.29 |
| bug-eagle | 0.29 |
| bug-insect | 0.05 |
| bug-swallow | 0.05 |
| bird-crow | 0.03 |
| bird-eagle | 0.03 |
| bird-insect | 0.35 |
| bird-swallow | 0.02 |
| crow-eagle | 0 |
| crow-insect | 0.55 |
| crow-swallow | 0.11 |
| eagle-insect | 0.55 |
| eagle-swallow | 0.11 |
| insect-swallow | 0.20 |

Table B.8: Cosine distance measures between all pairs of nouns from table B.6.

Then the clustering algorithm proceeds by calculating the distances between all the clusters existing at this second step: cluster ${ }_{1}=\{$ bee $\}$, cluster ${ }_{2}=\{b u g\}$, cluster $_{3}=$
 cluster $_{8}=\{$ crow, eagle $\} .{ }^{12}$ It turns out that the next closest pair is bee-insect (step 3), then bird-swallow (step 4), \{crow, eagle $\}$ - $\{$ bird, swallow $\}$ (step 5), bug- $\{$ bee, insect $\}$ (step 6) and finally \{crow, eagle, bird,swallow\}-\{bug, bee, insect\} (step 7).

| step number | cluster merged |
| ---: | :---: |
| 2 | $\{$ crow $\} \cup\{$ eagle $\}$ |
| 3 | $\{$ bee $\} \cup\{$ insect $\}$ |
| 4 | $\{$ bird $\} \cup\{$ swallow $\}$ |
| 5 | $\{$ crow,eagle $\cup\{$ bird,swallow $\}$ |
| 6 | $\{$ bug $\} \cup\{$ bee,insect $\}$ |
| 7 | \{crow,eagle,bird,swallow $\} \cup\{$ bug,bee,insect $\}$ |

Table B.9: Clusters obtained at each step from a single linkage clustering and cosine distance on data from table B.6.

Table B. 9 summarises all clustering steps, and figure B. 4 shows the output dendrogram. ${ }^{13}$ The $x$-axis in figure B. 4 displays the clustered objects, while the $y$-axis shows

[^76]

Figure B.4: Dendrogram from the clustering represented in table B.9.
the distances between the clusters, called links. For example, the distance between bird and swallow is between 0.020 and 0.025 .

Cutting the tree The division of the tree into clusters can be obtained by specifying the number of desired clusters, a distance threshold, or other criteria, such as an inconsistency coefficient. ${ }^{14}$ For example, if we wanted to cut the tree in figure B. 4 setting a distance threshold of 0.031 , we would get the following two clusters: \{crow,eagle, bird, swallow\} and \{bee, insect, bug\}.

## B. 6 Knowledge-free approaches in SP acquisition

In this section we will illustrate previous work in distributional methods for SP acquisition which were briefly described in section 5.2.2.

## B.6.1 Rooth et al. (1999a)

Rooth et al. (1999a) (see section 5.2.2.3) propose to find the best classes $c$ that can smooth the joint probability distribution $p(v s, n)$ of verb-slot-subcategorization frame triples $v s$ and nouns $n$. For example, ${ }^{15}$ the following verbs and slots
increase:subj:intrans, increase:obj:trans, fall:subj:intrans, decline:subj:intrans, ...

[^77]would form a class together with the nouns number, rate, price, amount. Both $v$ and $n$ are conditioned on $c$ :
$$
p(v s, n)=\sum_{c \in C} p(c, v s, n)=\sum_{c \in C} p(c) p(v s \mid c) p(n \mid c) .
$$

The best classes $c$ for the input data are found through the expectation-maximisation (EM) algorithm (Dempster and Rubin, 1977): this algorithm implements an iterative procedure: the first step (expectation) estimates an auxiliary function which produces the expected value of the log likelihood function given the data and the current fit of the parameter values; the second step (maximisation) maximises the auxiliary function as a function of the parameters.

## B.6.2 Erk (2007)

The system by Erk (2007) uses two corpora: a primary corpus and a generalization corpus. The primary corpus (over 59,000 sentences) is a semantic-role annotated corpus obtained from FrameNet (Baker et al., 1998); the generalization corpus (over 139,000 sentences) is the British National corpus (BNC), automatically dependencyparsed. Note that the primary corpus consists of manually semantically annotated data, where predicates correspond to semantic verb classes and arguments correspond to semantic roles; on the other hand, the generalization corpus is composed by syntactically annotated data, where the predicates correspond to verbs and arguments to their syntactic dependents. This way, Erk's system is able to calculate SPs for verb senses.

First, she extracts a set Seen $(r)$ of tuples (verbclass, relation, semantic role) from the primary corpus, for example (Morality_ evaluation, Evaluee, gamblers). ${ }^{16}$ She then calculates the SP for a word $w_{0}$ as:

$$
S\left(w_{0}\right)=\sum_{w \in \operatorname{Seen}(r)} \operatorname{sim}\left(w_{0}, w\right) \cdot w t_{r}(w)
$$

where $w t_{r}(w)$ stands for the weight of a word $w$ belonging to $\operatorname{Seen}(r)$ and $\operatorname{sim}\left(w_{0}, w\right)$ is the similarity between $w_{0}$ and $w$, calculated in the generalization corpus.

Erk tries several similarity metrics: Cosine, the Dice and Jaccard coefficients (see equations B. 5 and B.5), Hindle (1990)'s and Lin (1998)'s metrics. The weights $w$ range from the uniform distribution $w t_{r}(w)=1$ to frequency-based weights $w t_{r}(w)=$ $f r(w, r)$, to inverse document frequency. ${ }^{17}$

Erk compares her method to Resnik's model and to an EM clustering model, concluding that her method performs better, but has lower coverage with respect to the EM clustering.

## B.6.3 Bamman and Crane (2008)

Bamman and Crane (2008) propose the use of the log likelihood test for calculating association scores between single verbs and single nouns occurring as their arguments. Their input data is the parsed Perseus Digital Library corpus, consisting of 3.5 million words.

[^78]For each $(v, n)$ pair, they calculated

$$
\begin{array}{r}
\log \lambda=\log L\left(f r(v, n), f r(v), \frac{f r(n)}{N}\right)+\log L\left(f r(n)-f r(v, n), N-f r(v), \frac{f r(n)}{N}\right)+ \\
-\log L\left(f r(v, n), f r(v), \frac{f r(v, n)}{f r(v)}\right)-\log L\left(f r(n)-f r(v, n), N-f r(v), \frac{f r(n)-f r(v, n)}{N-f r(v)}\right),
\end{array}
$$

where $f r(v, n)$ is the frequency of $n$ as an argument of $v$ and $L(a, b, c)=c^{a}(1-c)^{b-a}$.
Given the high complexity of the LDT (various authors, genres and time periods), their system makes finer-grained distinctions between the specific usages in different authors, eras, and genres.

## B. 7 Data from Latin WordNet

In this section we provide some data extracted from LWN and relative to some lemmas used as examples in chapter 5.

## B.7.1 Synsets

In this section we list the WN synsets for the nouns culpa 'fault', esse 'being', fons 'source', forma 'form', homo 'man', pars 'part', solium 'throne', and terra 'earth', mentioned in section 5.3. We indicate the lemma, followed by an index for each synset, the domain assigned to the synset in MultiWordNet (MWN) and the gloss. Then, we represent the synset in LWN between two curly brackets, and the corresponding synset in the English Princeton WordNet (PWN). These data are taken from the MWN interface. ${ }^{18}$

## B.7.1.1 Culpa

culpa/1: (Factotum) [a wrong action attributable to bad judgment or ignorance or inattention; "the fault was all mine"] \{culpa, delictum, erratum, error, menda, mendum, peccatum, vitium $\}=\{$ mistake, error, fault $\}$
culpa/2: (Baseball, Play) [(baseball) a failure of a defensive player to make an out when normal play would have sufficed] $\{$ culpa $\}=\{$ error, misplay $\}$
culpa/3: (Table_Tennis, Tennis) [a serve that lands outside the prescribed area; "he served too many double faults"] \{culpa $\}=\{$ fault $\}$
culpa/4: (Factotum) [failure to act] \{cessatio, culpa, derelictio, derelictus, incuria, indiligentia, neglectus, neglegentia $\}=\{$ negligence, carelessness, neglect, nonperformance\}
culpa/5: (Psychological_Features) [the trait of neglecting responsibilities and lacking concern] \{cessatio, culpa, derelictio, derelictus, incuria, indiligentia, neglectus, neglegentia $\}=\{$ negligence, neglect, neglectfulness $\}$
culpa/6: (Psychological_Features) [ability or necessity to answer for or be responsible for one's conduct; "he holds a position of great responsibility"; "young children on a farm are often given responsibilities"] \{causa, culpa, procuratio $\}=\{$ responsibility, responsibleness\}
culpa/7: (Psychological_Features) [responsibility for a bad situation or event; "it was John's fault"] \{crimen, culpa, delictum, delinquentia, noxia, vitium $\}=\{$ fault $\}$

[^79]culpa/8: (Mechanics) [an imperfection in a device or machine; "if there are any defects you should send it back to the manufacturer"] \{crimen, culpa, delictum, labes, vitium $\}=\{$ defect, fault, flaw $\}$

## B.7.1.2 Esse

esse/1: (Factotum) [the state or fact of existing: "a point of view gradually coming into being"; "laws in existence for centuries"] \{esse, existentia $\}=\{$ being, beingness, existence\}

## B.7.1.3 Fons

fons 1: (Factotum) [a document (or organization) from which information is obtained; "the reporter had two sources for the story"] \{caput, fons, mater, origo $=$ \{source\}
fons $/ 2$ : (Factotum) [a point at which water issues forth] $\{$ fons $\}=\{$ spring $\}$
fons $/ 3$ : (Geology) [a natural flow of ground water] $\{$ fons $\}=\{$ spring, fountain, outflow, outpouring, natural_spring\}

## B.7.1.4 Forma

forma/1: (Factotum) [the act of appearing in public view; "the rookie made a brief appearance in the first period"; "it was Bernhardt's last appearance in America"] \{apparentia, conspectus, forma, imago, visus $\}=\{$ appearance $\}$
forma/2: (Factotum) [a predetermined set of movements in dancing or skating; "she made the best score on compulsory figures"] \{designatio, figura, forma, schema\} $=\{$ figure $\}$
forma/3: (Sociology) [pretending that something is the case in order to make a good impression; "they try to keep up appearances"; "that ceremony is just for show"] \{apparentia, aspectus, conspectus, facies, forma, imago, species, visus $\}=\{$ appearance, show $\}$
forma/4: (Factotum) [a representation of a bodily form (especially of a person); "he made a figure of Santa Claus"] \{designatio, figura, forma, schema $\}=\{$ figure $\}$
forma/5: (Factotum) [a container into which liquid is poured to create a given shape when it hardens] $\{$ forma $\}=\{$ mold, mould, cast $\}$
forma/6: (Factotum) [outward or visible aspect of a person or thing] \{apparentia, aspectus, conspectus, facies, forma, imago, species, visus $\}=\{$ appearance, visual_aspect $\}$
forma/7: (Factotum) [the impression produced by a person; "he cut a fine figure"; "a heroic figure"] \{designatio, figura, forma, schema\} $=\{$ figure $\}$
forma/8: (Factotum) [the visual appearance of something or someone; "the delicate cast of his features"] \{facies, figura, forma, formula, schema $\}=\{$ form, shape, cast $\}$
forma/9: (Factotum) [the qualities that give pleasure to the senses] \{decus, facies, figura, forma, pulchritudo, species $\}=\{$ beauty $\}$
forma/10: (Factotum) [any spatial attribute (especially as defined by outline); "he could barely make out their shapes through the smoke"] \{facies, figura, forma, formula, schema $\}=\{$ shape, form, configuration, contour $\}$
forma/11: (Factotum) [an ability to perform well; "he was at the top of his form"; "the team was off form last night"] \{figura, forma, schema\} $=\{$ form $\}$
forma/12: (Factotum) [a category of things distinguished by some common characteristic or quality; "sculpture is a form of art"; "what kinds of desserts are there?"] $\{$ figura, forma, genus, species, typus $\}=\{$ kind, sort, form, variety $\}$
forma/13: (Factotum) [an outstanding example of its kind; "his roses were beauties"; "when I make a mistake it's a beaut] \{decus, facies, figura, forma, pulchritudo, species $\}=\{$ beauty, beaut $\}$
forma/14: (Factotum) [a unitary percept having structure and coherence that is the object of attention and that stands out against a ground] \{figura, forma\} $=\{$ figure $\}$
forma/15: (Factotum) [a model considered worthy of imitation; "the American constitution has provided a pattern for many republics"] \{archetypum, exemplar, exemplum, forma, formula, typus $\}=\{$ pattern $\}$
forma/16: (Grammar) [the phonological or orthographic sound or appearance of a word; "the inflected forms of a word can be represented by a stem and a list of inflections to be attached"] \{figura, forma, schema $\}=\{$ form, word_form $\}$
forma/17: (Art, Literature) [an arrangement of the elements in a composition or discourse; "the essay was in the form of a dialogue"; "he first skethes the plot in outline form"] \{figura, forma, schema $\}=\{$ form $\}$
forma/18: (Literature) [language used in a figurative or nonliteral sense] \{figura, forma, schema $\}=$ \{trope, figure_of_speech, figure, image $\}$
forma/19: (Factotum) [the event of coming into sight] \{apparentia, aspectus, conspectus, forma, imago, species, visus $\}=\{$ appearance $\}$
forma/20: (Person) [a very attractive or seductive looking woman] \{decus, facies, figura, forma, pulchritudo, species $\}=\{$ smasher, stunner, knockout, beauty, sweetheart, peach, lulu, looker, mantrap, dish\}
forma/21: (Factotum) [a combination of points and lines and planes that form a visible palpable shape] \{designatio, figura, forma, schema $\}=\{$ figure $\}$

## B.7.1.5 Homo

homo/1: (Person) [a human being; "there was too much for one person to do"] $\{$ anima, animus, caput, corpus, homo, pectus, spiritus $\}=\{$ person, individual, someone, somebody, mortal, human, soul\}
homo/2: (Biology, Person) [any living or extinct member of the family Hominidae] $\{$ homo, vir $\}=\{$ homo, man, human_being, human $\}$
homo/3: (Factotum) [a person's body (usually including their clothing); "a weapon was hidden on his person"] \{caput, corpus, homo $\}=\{$ person $\}$
homo/4: (Grammar) [a grammatical category of pronouns and verb forms; "stop talking about yourself in the third person"] \{caput, corpus, homo $\}=\{$ person $\}$
homo/5: (Factotum) [all of the inhabitants of the earth; "all the world loves a lover"] $=\{$ caelum, homo, mundus, vir $\}=\{$ world, human_race, humanity, humankind, human_beings, humans, mankind, man\}
homo/6: (Person) [an adult male person (as opposed to a woman); "there were two women and six men on the bus"] \{homo, vir $\}=\{$ man, adult_male $\}$
homo/7: (Person) [(informal) a male person who plays a significant role (husband or lover or boyfriend) in the life of a particular woman; "she takes good care of her man"] $\{$ homo, vir $\}=\{\operatorname{man}\}$
homo/8: (Person) [an adult male person who has a manly character (virile and courageous competent); "the army will make a man of you"] \{homo, vir $\}=\{\operatorname{man}\}$
homo/9: (Person) [a male subordinate; "the chief stationed two men outside the building"; "he awaited word from his man in Havana"] \{homo, vir $\}=\{\operatorname{man}\}$
homo/10: (Military) [someone who serves in the armed forces; "two men stood sentry duty"] \{homo, vir $\}=\{$ serviceman, military_man, man, military_personnel $\}$

## B.7.1.6 Pars

pars /1: (Factotum) [a job in an organization or hierarchy; "he occupied a post in the treasury"] \{causa, cura, dignitas, munus, officium, pars, provincia $\}=\{$ position, post, berth, slot, office, spot, place, situation $\}$
pars/2: (Factotum) [the actions and activities assigned to or required or expected of a person or group: "the function of a teacher"; "the government must do its part" or "play its role" or "do its duty"] \{cura, functio, munus, officium, pars, provincia\} $=$ \{function, office, part, role\}
pars/3: (Factotum) [any one of a number of individual efforts in a common endeavor: "I am proud of my contribution to the team's success"; "they all did their share of the work"] \{intributio, pars, portio $\}=\{$ contribution, part, share $\}$
pars/4: (Religion) [a religious rite or service prescribed by ecclesiastical authorities; "the offices of the mass"] \{cura, ministerium, munus, officium, pars, provincia\} $=$ \{office\}
pars /5: (Factotum) [the act of managing something; "he was given overall management of the program"; "is the direction of the economy a function of government?"] \{administratio, curatio, cursus, dispensatio, gubernatio, moderamen, pars, procuratio, regio $\}=\{$ management, direction, managing $\}$
pars/6: (Buildings) [where professional or clerical duties are performed; "he rented an office in the new building"] \{ministerium, officium, pars $\}=\{$ office $\}$
pars $/ 7$ : (Factotum) [what something is used for; "the function of an auger is to bore holes"; "ballet is beautiful but what use is it?] \{consummatio, functio, munus, pars $\}=$ \{function, purpose, role, use\}
pars 18: (Anatomy) [a part of an animal that has a special function or is supplied by a given artery or nerve; "in the abdominal region"] \{arvum, cardo, orbis, pars, plaga, populus, regio, terra, tractus $\}=\{$ area, region $\}$
pars /9: (Theatre) [an actor's portrayal of someone in a play; "she played the part of Desdemona"] \{pars, persona, portio $\}=\{$ character, role, theatrical_role, part, persona $\}$ pars $/ 10$ : (Music) [the melody carried by a particular voice or instrument in polyphonic music; "he tried to sing the tenor part"] \{pars, portio, vox $\}=\{$ part, voice $\}$
pars/11: (Gastronomy) [an individual portion of food or drink: "the helpings were all small"; "his portion was larger than hers"; "there's enough for two servings each"] \{pars, portio $\}=\{$ helping, portion, serving $\}$
pars/12: (Factotum) [a dissenting clique] \{factio, grex, pars $\}=\{$ faction, sect $\}$
pars /13: (History, Politics) [an organization to gain political power; "in 1992 Perot tried to organize a third party at the national level"] \{factio, pars $\}=\{$ party, political_party\}
pars /14: (Military, Play, Politics) [one of two or more contesting groups (in games or war or politics); "the Confederate side was prepared to attack"] \{latus, pars $\}=\{$ side $\}$
pars 15 : (Factotum) [a large indefinite location on the surface of the Earth; "penguins inhabit the polar regions"] \{arvum, orbis, pars, plaga, populus, regio, terra, tractus $\}=\{$ region $\}$
pars $/ 16$ : (Law) [a person involved in legal proceedings; "the party of the first part"] $\{$ pars $\}=\{$ party $\}$
pars 17 : (Mathematics) [a mathematical relation such that each element of one set is associated with at least one element of another set] \{functio, munus, pars\} $=$ \{function, mathematical_function\}

## B.7.1.7 Solium

solium 1: (Furniture) [the chair of state of a monarch, bishop, etc.] \{solium, thronus $\}=\{$ throne $\}$

## B.7.1.8 Terra

terra/1: (Electricity) [a connection between an electrical device and the earth (which is a zero voltage)] \{humus, tellus, terra $\}=\{$ ground, earth $\}$
terra/2: (Anatomy) [a part of an animal that has a special function or is supplied by a given artery or nerve; "in the abdominal region"] \{arvum, cardo, orbis, pars, plaga, populus, regio, terra, tractus $\}=\{$ area, region $\}$
terra/3: (Politics) [the people of a nation or country or a community of persons bound by a common heritage; "a nation of Catholics"; "the whole country worshipped him""] \{gens, humus, natio, populus, solum, tellus, terra $\}=\{$ nation, nationality, land, country, a_people\}
terra/4: (Politics) [a politically organized body of people under a single government; "the state has elected a new president"] \{adfectus, civitas, gens, humus, populus, solum, tellus, terra $\}=\{$ state, nation, country, land, commonwealth, res_publica, body_politic $\}$
terra/5: (Geography) [a particular geographical region of indefinite boundary (usually serving some special purpose or distinguished by its people or culture or geography); "it was a mountainous area"; "Bible country"] \{area, arvum, cardo, humus, mensura, regio, spatium, tellus, terra $\}=\{$ area, country $\}$
terra/6: (Administration, Geography) [the territory occupied by a nation; "he returned to the land of his birth"; "he visited several European countries"] \{adfectus, arvum, civitas, gens, gleba, humus, natio, populus, solum, tellus, terra $\}=\{$ country, state, land, nation $\}$
terra/7: (Religion) [the abode of mortals (as contrasted with heaven or hell); "it was hell on earth"] \{humus, tellus, terra $\}=\{$ Earth $\}$
terra/8: (Factotum) [a large indefinite location on the surface of the Earth; "penguins inhabit the polar regions""] \{arvum, orbis, pars, plaga, populus, regio, terra, tractus $\}=\{$ region $\}$
terra/9: (Geography) [the solid part of the earth's surface; "the plane turned away from the sea and moved back over land""; "the earth shook for several minutes"; "he dropped the logs on the ground"] \{arvum, gleba, humus, tellus, terra $\}=\{$ land, dry_land, earth, ground, solid_ground, terra_firma\}
terra/10: (Geography) [what plants grow in (especially with reference to its quality or use); "the land had never been plowed"; "good agricultural soil"] \{ager, arvum, gleba, humus, solum, tellus, terra $\}=\{$ land, ground, soil $\}$
terra/11: (Astronomy) [the 3rd planet from the sun; the planet on which we live; "the Earth moves around the sun"; "he sailed around the world"] \{caelum, globus, humus, mundus, sphaera, tellus, terra $\}=\{$ Earth, world, globe $\}$
terra/12: (Economy) [extensive landed property (especially in the country) retained by the owner for his own use; "the family owned a large estate on Long Island"] \{ager, fundus, gleba, humus, praedium, solum, tellus, terra, villa $\}=\{$ estate, land, landed_estate, acres, demesne\}
terra/13: (Geology) [the loose soft material that makes up a large part of the surface of the land surface; "They dug into the earth outside the church"] \{humus, solum, tellus, terra $\}=\{$ earth, ground $\}$
terra/14: (Philosophy) [(archaic) once thought to be one of four elements composing the universe] $\{$ humus, tellus, terra $\}=\{$ earth $\}$

## B.7.1.9 Mundus

mundus $/ 1$ : (Factotum) [something used to beautify] \{mundus, ornamentum $\}=$ \{decoration, ornament, ornamentation\}
mundus/2: (Psychology) [the concerns of the world as distinguished from heaven and the afterlife; "they consider the church to be independent of the world"] \{caelum, mundus $\}=\{$ worldly_concern, earthly_concern, world, earth $\}$
mundus/3: (Factotum) [all of the inhabitants of the earth; "all the world loves a lover"] \{caelum, homo, mundus, vir\} = \{world, human_race, humanity, humankind, human_beings, humans, mankind, man\}
mundus/4: (Astronomy) [the apparent surface of the imaginary sphere on which celestial bodies appear to be projected] \{aethra, caelum, mundus $\}=\{$ celestial_sphere, sphere, empyrean, firmament, heavens, vault_of_heaven, welkin\}
mundus/5: (Astronomy, Physics) [everything that exists anywhere; "they study the evolution of the universe"; "the biggest tree in existence"] \{caelum, cosmos, mundus, universitas $\}=\{$ universe, existence, nature, creation, world, cosmos, macrocosm $\}$
mundus/6: (Astronomy) [the 3rd planet from the sun; the planet on which we live; "the Earth moves around the sun"; "he sailed around the world"] \{caelum, globus, humus, mundus, sphaera, tellus, terra $\}=\{$ Earth, world, globe $\}$

## B.7.2 Examples of semantic properties

In this section we list the semantic properties for homo 'man', terra 'earth' and mundus 'world'.

## B.7.2.1 Homo

\{adsecla, adsectator, anatomy, anima, animus, being, bod, build, caelum, caput, causa, chassis, classis, cognatio, corpus, entity, familia, figure, flesh, focus, form, frame, genus, grammatical_category, grex, hominid, homo, human_body, life_form, living_thing, male, male_person, material_body, mundus, natio, ordo, organism, pectus, physical_body, physique, res, shape, skilled_worker, soma, something, spiritus, subordinate, subsidiary, syntactic_category, trained_worker, underling, vir\}

## B.7.2.2 Terra

\{administrative_district, administrative_division, affectus, ager, area, arvum, body_part, caelum, cardo, census, civitas, classis, complexio, detentatio, detentio, elementum, fundus, gens, glaeba, globus, humus, instrumentality, instrumentation, location, materia, mensura, mundus, natio, natural_object, object, orbis, organisation, pars, physical_object, plaga, planeta, political_unit, populus, possessio, praedium, real_estate, real_property, realty, regio, res, solum, spatium, sphaera, stella, substantia, tellus, terra, terrestrial_planet, territorial_division, toparchia, tractus, villa\}

## B.7.2.3 Mundus

\{adfectus, aethra, artefact, artifact, astrum, caelum, circumscriptio, civitas, classis, cognitive_state, commodum, concern, confine, cosmos, curiosity, entity, extremitas, fenus, finis, globus, homo, humus, location, meta, modus, mundus, natural_object, object, ornamentum, part, physical_object, planeta, region, something, sphaera, state_of_mind, stella, surface, tellus, termen, terminus, terra, terrestrial_planet, universitas, usura, vir, wonder\}

## B. 8 Formulae for the probabilistic model

In this section we will give the formulae for our SP acquisition system.

## B.8.1 Clustering of frames

Here we give the details of the formulae involved in the clustering of frames introduced in 5.3.2.1.

Let $F$ be a frame, defined according to the criteria given in 5.3.1.1. A construction $K$ is assigned to $F$ by the clustering algorithm if $K$ maximises the probability $P(k \mid F)$ over all constructions $k$, including the baseline $k_{0}=\{F\}$ :

$$
K=\arg \max _{k} P(k \mid F)
$$

After Bayes' theorem, since $P(F)$ is constant for all $k$ :

$$
K=\arg \max _{k} P(k) P(F \mid k)
$$

- $P(k)$ is calculated as the number $n_{k}$ of frames contained in $k$ divided by $N+1$, where $N$ is the total number of frames:

$$
P(k)=\frac{n_{k}}{N+1}
$$

Hence, the prior probability of a new construction is $P\left(k_{0}\right)=\frac{1}{N+1}$.

- If we assume that the frame features are independent, $P(F \mid k)$ is the product of $P_{i}\left(\right.$ feature $\left._{i}(F) \mid k\right)$ for $i=1,2,3$ :

$$
\begin{equation*}
P(F \mid k)=\prod_{i=1,2,3} P_{i}\left(\text { feature }_{i}(F) \mid k\right) \tag{B.7}
\end{equation*}
$$

where $P_{i}\left(\right.$ feature $\left._{i}(F) \mid k\right)$ is the probability that the $i^{\text {th }}$ feature displays the same value in $F$ and in $k$, that is feature ${ }_{i}(F)$.

- Feature $_{1}$. We estimated $P\left(\right.$ feature $\left._{1}(F) \mid k\right)$ by using the following smoothed MLE formula:

$$
P\left(\text { feature }_{1}(F) \mid k\right)=\frac{\left(\sum_{h \in k} \text { synt_score }(h, F)\right)+\lambda}{\left.n_{k}\right)+\lambda \alpha_{1}}
$$

where $\operatorname{synt}$ _score $(h, F)=\frac{|S C S(h) \cap S C S(F)|}{|S C S(F)|}$ (syntactic score) is the number of syntactic slots shared by $h$ and $F$ over the number of slots in $F ; n_{k}$ is the number of frames in $k$. The smoothing factor $\alpha_{1}$ is set to an estimate of the number of possible values that feature ${ }_{1}$ can have. Thus, $\alpha_{1}=1537$, the number of all slots realised in the dataset; we set $\lambda=10^{-7}$. See Alishahi (2008, p. 49) for a discussion on how to set $\lambda$.

- Feature $_{2}$. Let $\operatorname{nslots}(F)$ be the number of argument slots in $F$; for each such slot a, $P\left(\right.$ feature $\left._{2}(F) \mid k\right)$ is

$$
\begin{equation*}
P\left(\text { feature }_{2}(F) \mid k\right)=\frac{\sum_{a}\left[\sum_{h \in k} \text { semantic_score }_{a}(h, F)\right]+\lambda}{n_{k} \cdot \operatorname{nslots}(F)+\lambda \alpha_{2}} \tag{B.8}
\end{equation*}
$$

where semantic_score $a_{a}(h, F)=\frac{|S(h) \cap S(F)|}{|S(F)|}$ (semantic score) accounts for the degree of overlap between the semantic properties $S(h)$ of $h$ and the semantic properties $S(F)$ of $F$ (for argument $a$ ). $\alpha_{2}$ is the number of semantic properties of all fillers in the treebanks (697)

- Feature $_{3}$ : the probability of displaying the same value in $F$ and in $k$, i. e. feature ${ }_{3}$, is

$$
\begin{equation*}
P\left(\text { feature }_{3}(F) \mid k\right)=\frac{\sum_{h \in k} \text { synset_score }^{( }(h, F)}{n_{k}} \tag{B.9}
\end{equation*}
$$

where synset_score $(h, F)=\frac{\mid \operatorname{Synsets}(\text { verb }(h)) \text { ) } \operatorname{Synsets}(\operatorname{ver} b(F)) \mid}{|\operatorname{Synsets}(v e r b(F))|}$ (synset score) calculates the degree of 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 $F$.

## B.8.2 SP probability distribution

The probability of a semantic property $s$ in an argument position $a$ for a verb $v$ is calculated as:

$$
P_{a}(s \mid v)=\sum_{k} P_{a}(s, k \mid v) \propto \sum_{k} \frac{P(k, v) P_{a}(s \mid k, v)}{P(v)}
$$

- $P(k, v)$ is estimated as

$$
P(k, v)=\frac{n_{k} \cdot \operatorname{freq}(k, v)}{\sum_{k^{\prime}} n_{k} \operatorname{freq}\left(k^{\prime}, v\right)}
$$

where $n_{k}$ is the number of frames contained in $k$ whose verbs share a synset with $v$.

- We set $S P_{-}$similarity_score $a_{a}(h, v, s)$ as:

$$
\begin{equation*}
\operatorname{freq}(h) * \text { synset_score }(v, \operatorname{verbs}(h)) * \text { semantic_score }(s, h) \tag{B.10}
\end{equation*}
$$

with

$$
\begin{equation*}
\operatorname{synset} \text { _score }(v, \operatorname{verbs}(h))=\frac{|\operatorname{Synsets}(v) \cap \operatorname{Synsets}(\operatorname{verbs}(h))|}{|\operatorname{Synsets}(v) \cup \operatorname{Synsets}(\operatorname{verbs}(h))|} \tag{B.11}
\end{equation*}
$$

and

$$
\begin{equation*}
\operatorname{semantic\_ score}(s, h)=\frac{\sum_{f} \operatorname{semantic\_ score}(s, S(f))}{N_{\text {fillers }(h, a)}}=\frac{\sum_{f} \frac{|s \cap S(f)|}{|s \cup S(f)|}}{N_{\text {fillers }(h, a)}} \tag{B.12}
\end{equation*}
$$

$S(f)$ in (B.12) is the set of semantic properties of each filler $f$ of $\operatorname{verbs}(h)$ for $a$ and $N_{\text {fillers }(h, a)}$ is the number of fillers $f$ for $a$ in $h$ : this way the value in (B.10) ranges between 0 (when synset_score $(v, \operatorname{verbs}(h)=0$ and/or semantic_score $(s, h)=0)$ and $\operatorname{freq}(h)$ (when $\operatorname{synset\_ score}\left(v, \operatorname{verbs}(h)=1\right.$ and $\left.\operatorname{semantic\_ score}(s, h)=1\right)$.

- We estimate $P_{a}(s \mid k, v)$ by normalising over all $k$

$$
\begin{equation*}
\frac{\left(\sum_{h} S P_{-} \text {similarity_score } a(h, v, s)\right)+\lambda \alpha_{s}}{\left(\sum_{h} \text { freq }(h, v)\right)+\lambda \alpha} \tag{B.13}
\end{equation*}
$$

where $h$ is summed over all frames in $k$ whose verbs share the same synset with $v$. $\lambda, \alpha$ and $\alpha_{s}$ are smoothing factors: $\alpha_{s}$ is set as the relative frequency of $s$ among all fillers and $\alpha$ is the sum of all $\alpha_{s}$. We set $\lambda=10^{-7}$, as in the clustering step. See Alishahi (2008, p. 128) for a discussion on how to set $\lambda$.

Evaluation In the evaluation phase, we were interested in calculating the probability of a specific noun $n$ occurring in an argument position $a$ for a verb $v$. Therefore, we changed formula (B.10) into

$$
\begin{equation*}
f r e q(h) * \text { synset_score }(v, \operatorname{verbs}(h)) * \frac{\sum_{f} \operatorname{semantic\_ score}(S(n), S(f))}{N_{\text {fillers }(h, a)}} \tag{B.14}
\end{equation*}
$$

where the semantic score is calculated between the set of semantic properties of $n$ and those of each filler $f$ for $a$ in $h$. The other quantities remain the same as in section 5.3.2.3.

Cosine distance As explained in section 5.4.3, we evaluated a distributional version of our algorithm, where the similarity scores were calculated on the basis of the cosine distance. For each noun lemma $n$ in the input parsed corpora, we calculated $P_{a}(n \mid k, v)$ as

$$
\begin{equation*}
\text { freq }(h) * \text { cosine_distance }(v, \text { verbs }(h)) * \frac{\sum_{f} \operatorname{cosine\_ distance~}(n, f)}{N_{\text {fillers }(h, a)}} \tag{B.15}
\end{equation*}
$$

where the first cosine distance is calculated between the vector representations of $v$ and the verb in $h$; the second cosine distance is calculated between the noun $n$ and each filler $f$ for $a$ in $h$. Correspondingly, $P_{a}(n \mid k, v)$ is calculated as

$$
\begin{equation*}
\left.\left.\frac{\left(\sum_{h} S P \_\right. \text {similarity_score }}{a}(h, v, n)\right)+\lambda \alpha_{n}\right) \tag{B.16}
\end{equation*}
$$

The smoothing factor $\alpha_{n}$ is set as the relative frequency of $n$ among all fillers in the parsed corpora and $\beta$ is the sum of all $\alpha_{n}$.

## B. 9 Formulae for the evaluation of the SP system

In this section we give some details for the evaluation of our SP acquisition system described in section 5.4. Section B.9.0.1 illustrates the measures of statistical despersion we referred to while analysing the shape of the SP distributions, while section B.9.0.2 covers the measures of central tendency we used.

## B.9.0.1 Statistical dispersion

In order to assess how skewed the SP probability distribution given by our WN-model is, we calculated the values of two measures of statistical dispersion on the evaluation set containing the distributions of 40 random verbs extracted from our corpora.

First, we calculated the Interquartile Range (IQR), defined as the difference between the third and first quartiles. By its definition, the IQR is less sensitive to outliers than
the usual range, which measures the difference the maximum and the minimum values of a set.

Second, we calculated the Median Absolute Deviation (MAD), by taking the median of the absolute deviations from the distributions' medians: ${ }^{19}$

$$
M A D(v)=\operatorname{median}\left(\left|P_{a}\left(n_{i} \mid v\right)-\operatorname{median}\left(P_{a}(\cdot \mid v)\right)\right|\right)
$$

For instance, the median of the set $(1,2,4,5,5,9)$ is 4.5 and the absolute deviations from the median are $(3.5,2.5,0.5,0.5,0.5,4.5)$; the median of the latter set is 2.22 , which is the MAD of the original set. Table B. 10 shows the IQR and MAD values for all 40 verbs in the evaluation set.

## B.9.0.2 Central tendency

While evaluating our system for SP acquisition, we analysed the probability distributions obtained for 40 random verbs, for which an example is given in figure 5.9 on page 110. We considered the 40 functions defined for each verb $v$

$$
P_{a}(\cdot \mid v): n \longmapsto P_{a}(n \mid v) \quad \forall n \in N
$$

These functions assign the SP probability to each noun $n$ belonging to the set $N$ of the 4724 LWN leaf nodes.

There are several ways to synthetically describe the typical value of a distribution. First, we calculated the mean of the values of the probability distribution:

$$
\frac{\sum_{n \in N} P_{a}(n \mid v)}{4724}
$$

For example, the mean for dico 'say' with the slot 'A_Obj[acc]' is 0.00021 .
Second, we calculated the median: we sorted the values $P_{a}(n \mid v)$ in descending order and picked the noun in the middle of this sequence, or the mean of the two middle points in case of an even number of observations. The set $N$ is thus split in two subsets: half the nouns have a probability lower than the median and half have a higher probability. The median for dico is 0.00013 .

Third, we calculated the values at the third quartile. We divided $N$ in four equalsized parts (quartiles); this way, $25 \%$ of the nouns have a probability higher than the value at the third quartile. The value of the third quartile for dico is 0.0003 . Figure B. 5 shows the mean, median and third quartile values for the distribution of the direct objects of dico, while table B. 11 shows the mean, median and value of third quartile for each of the 40 verbs we used for evaluation.

For each verb-slot pair, we calculated the value $P_{a}(f \mid v)$ predicted by the model for the random filler $f$ and compared it with the measures of central tendency just illustrated. Table B. 12 records each of the 40 verbs in the evaluation set, together with its frequency, its respective random filler, its SP probability and whether or not this probability is higher than each of the three measures of central tendency.

[^80]

Figure B.5: SP probability distribution for the objects of dico 'say'.

|  | IQR | MAD |
| ---: | ---: | ---: |
| accido2 | $2.50 \mathrm{e}-04$ | $1.89 \mathrm{e}-09$ |
| agito | $3.11 \mathrm{e}-04$ | $1.81 \mathrm{e}-04$ |
| ascendo | $3.15 \mathrm{e}-04$ | $1.56 \mathrm{e}-04$ |
| canto | $1.18 \mathrm{e}-04$ | $1.94 \mathrm{e}-06$ |
| cognosco | $2.64 \mathrm{e}-04$ | $1.59 \mathrm{e}-04$ |
| comparo | $3.29 \mathrm{e}-04$ | $1.86 \mathrm{e}-04$ |
| constituo | $3.12 \mathrm{e}-04$ | $1.75 \mathrm{e}-04$ |
| convenio | $2.91 \mathrm{e}-04$ | $2.07 \mathrm{e}-04$ |
| cupio | $2.88 \mathrm{e}-04$ | $1.87 \mathrm{e}-04$ |
| determino | $1.99 \mathrm{e}-09$ | $1.16 \mathrm{e}-09$ |
| dico2 | $2.63 \mathrm{e}-04$ | $1.92 \mathrm{e}-04$ |
| dimitto | $1.24 \mathrm{e}-04$ | $1.01 \mathrm{e}-09$ |
| do | $2.66 \mathrm{e}-04$ | $1.96 \mathrm{e}-04$ |
| facio | $2.23 \mathrm{e}-04$ | $1.65 \mathrm{e}-04$ |
| fero | $2.87 \mathrm{e}-04$ | $1.72 \mathrm{e}-04$ |
| induco | $3.05 \mathrm{e}-04$ | $1.82 \mathrm{e}-04$ |
| intellego | $2.82 \mathrm{e}-09$ | $1.86 \mathrm{e}-09$ |
| intro | $2.62 \mathrm{e}-04$ | $9.82 \mathrm{e}-10$ |
| invenio | $2.93 \mathrm{e}-04$ | $2.04 \mathrm{e}-04$ |
| maneo | $2.68 \mathrm{e}-04$ | $1.95 \mathrm{e}-04$ |
| moveo | $3.08 \mathrm{e}-04$ | $1.89 \mathrm{e}-04$ |
| multiplico | $2.56 \mathrm{e}-04$ | $1.90 \mathrm{e}-04$ |
| nego | $2.96 \mathrm{e}-04$ | $1.75 \mathrm{e}-04$ |
| paro | $2.89 \mathrm{e}-04$ | $1.36 \mathrm{e}-04$ |
| pateo | $2.42 \mathrm{e}-04$ | $1.82 \mathrm{e}-04$ |
| pono | $2.59 \mathrm{e}-04$ | $1.78 \mathrm{e}-04$ |
| praecedo | $2.62 \mathrm{e}-04$ | $1.94 \mathrm{e}-04$ |
| praedico | $2.35 \mathrm{e}-04$ | $7.38 \mathrm{e}-10$ |
| profero | $1.24 \mathrm{e}-04$ | $7.17 \mathrm{e}-10$ |
| recipio | $2.34 \mathrm{e}-04$ | $1.71 \mathrm{e}-04$ |
| refero | $4.29 \mathrm{e}-09$ | $2.82 \mathrm{e}-09$ |
| resulto | $3.18 \mathrm{e}-04$ | $1.86 \mathrm{e}-04$ |
| scio | $2.05 \mathrm{e}-04$ | $1.52 \mathrm{e}-10$ |
| sequor | $2.76 \mathrm{e}-04$ | $1.62 \mathrm{e}-04$ |
| subeo | $2.64 \mathrm{e}-04$ | $2.36 \mathrm{e}-09$ |
| sum | $2.22 \mathrm{e}-04$ | $1.44 \mathrm{e}-04$ |
| tento | $3.07 \mathrm{e}-04$ | $2.02 \mathrm{e}-04$ |
| timeo | $1.76 \mathrm{e}-04$ | $2.25 \mathrm{e}-09$ |
| venio | $2.62 \mathrm{e}-04$ | $1.24 \mathrm{e}-04$ |
| volo | $2.89 \mathrm{e}-04$ | $1.65 \mathrm{e}-04$ |
|  | 2 |  |

Table B.10: Interquartile Range (IQR) and Median Absolute Deviation (MAD) of the SP probability distributions given by our WN-based system, for each verb in the evaluation set. 'Accido2' and 'dico2' refer to the 'fall' and 'say' senses of the two verbs, respectively.

| verb | mean | median | third quartile |
| ---: | ---: | ---: | ---: |
| accido2 | 0.00021 | $1.273 \mathrm{e}-09$ | $2.496 \mathrm{e}-04$ |
| agito | 0.00021 | $1.218 \mathrm{e}-04$ | $3.114 \mathrm{e}-04$ |
| ascendo | 0.00021 | $1.054 \mathrm{e}-04$ | $3.150 \mathrm{e}-04$ |
| canto | 0.00021 | $1.310 \mathrm{e}-06$ | $1.180 \mathrm{e}-04$ |
| cognosco | 0.00021 | $1.074 \mathrm{e}-04$ | $2.635 \mathrm{e}-04$ |
| comparo | 0.00021 | $1.257 \mathrm{e}-04$ | $3.286 \mathrm{e}-04$ |
| constituo | 0.00021 | $1.181 \mathrm{e}-04$ | $3.120 \mathrm{e}-04$ |
| convenio | 0.00021 | $1.398 \mathrm{e}-04$ | $3.157 \mathrm{e}-04$ |
| cupio | 0.00021 | $1.264 \mathrm{e}-04$ | $2.88 \mathrm{e}-04$ |
| determino | 0.00021 | $8.575 \mathrm{e}-10$ | $2.118 \mathrm{e}-09$ |
| dico2 | 0.00021 | $1.292 \mathrm{e}-04$ | $2.897 \mathrm{e}-04$ |
| dimitto | 0.00021 | $6.903 \mathrm{e}-10$ | $1.244 \mathrm{e}-04$ |
| do | 0.00021 | $1.325 \mathrm{e}-04$ | $2.930 \mathrm{e}-04$ |
| facio | 0.00021 | $1.489 \mathrm{e}-04$ | $2.790 \mathrm{e}-04$ |
| fero | 0.00021 | $1.162 \mathrm{e}-04$ | $3.156 \mathrm{e}-04$ |
| induco | 0.00021 | $1.229 \mathrm{e}-04$ | $3.049 \mathrm{e}-04$ |
| intellego | 0.00021 | $1.401 \mathrm{e}-09$ | $3.049 \mathrm{e}-09$ |
| intro | 0.00021 | $6.726 \mathrm{e}-10$ | $2.624 \mathrm{e}-04$ |
| invenio | 0.00021 | $1.376 \mathrm{e}-04$ | $3.280 \mathrm{e}-04$ |
| maneo | 0.00021 | $1.316 \mathrm{e}-04$ | $2.951 \mathrm{e}-04$ |
| moveo | 0.00021 | $1.272 \mathrm{e}-04$ | $3.085 \mathrm{e}-04$ |
| multiplico | 0.00021 | $1.282 \mathrm{e}-04$ | $2.565 \mathrm{e}-04$ |
| nego | 0.00021 | $1.178 \mathrm{e}-04$ | $2.958 \mathrm{e}-04$ |
| paro | 0.00021 | $9.173 \mathrm{e}-05$ | $2.918 \mathrm{e}-04$ |
| pateo | 0.00021 | $1.230 \mathrm{e}-04$ | $2.816 \mathrm{e}-04$ |
| pono | 0.00021 | $1.200 \mathrm{e}-04$ | $2.827 \mathrm{e}-04$ |
| praecedo | 0.00021 | $1.309 \mathrm{e}-04$ | $2.621 \mathrm{e}-04$ |
| praedico | 0.00021 | $5.183 \mathrm{e}-10$ | $2.348 \mathrm{e}-04$ |
| profero | 0.00021 | $4.892 \mathrm{e}-10$ | $1.238 \mathrm{e}-04$ |
| recipio | 0.00021 | $1.152 \mathrm{e}-04$ | $2.623 \mathrm{e}-04$ |
| refero | 0.00021 | $2.126 \mathrm{e}-09$ | $4.632 \mathrm{e}-09$ |
| resulto | 0.00021 | $1.256 \mathrm{e}-04$ | $3.180 \mathrm{e}-04$ |
| scio | 0.00021 | $1.042 \mathrm{e}-10$ | $2.049 \mathrm{e}-04$ |
| sequor | 0.00021 | $1.090 \mathrm{e}-04$ | $2.762 \mathrm{e}-04$ |
| subeo | 0.00021 | $1.594 \mathrm{e}-09$ | $2.642 \mathrm{e}-04$ |
| sum | 0.00021 | $1.523 \mathrm{e}-04$ | $2.773 \mathrm{e}-04$ |
| tento | 0.00021 | $1.360 \mathrm{e}-04$ | $3.073 \mathrm{e}-04$ |
| timeo | 0.00021 | $1.519 \mathrm{e}-09$ | $1.760 \mathrm{e}-04$ |
| venio | 0.00021 | $8.343 \mathrm{e}-05$ | $2.623 \mathrm{e}-04$ |
| volo | 0.00021 | $1.114 \mathrm{e}-04$ | $2.887 \mathrm{e}-04$ |
|  |  |  |  |

Table B.11: Mean, median and third quartile of the SP probability distributions given by our WN-based system, for each verb in the evaluation set. 'Accido2' and 'dico2' refer to the 'fall' and 'say' senses of the two verbs, respectively.

Table B.12: Frequency and random filler for each verb in the evaluation set, followed by a binary value indicating whether the SP probability of the random filler is higher than the mean, median and third quartile, respectively. 'Accido2' and 'dico2' refer to the 'fall' and 'say' senses of the two verbs, respectively.

The last row of table B. 12 shows the total number of pairs for which the SP probability of the random filler was higher than the measure of random tendency: 28 (mean), 36 (median) and 24 (third quartile). We now want to answer the following question: are these proportions significantly higher than we would expect by chance?

Binomial test To answer this question we performed a binomial test in R ( R Development Core Team, 2008), which tested the null hypothesis that the successes (probability higher than the measure of central tendency) and the failures (lower probability) are equally likely.

For the 28 successes (mean), the test gave the $95 \%$ confidence interval [0.53, 0.83] and estimated the probability of success as 0.7 . For the 36 successes (median), the $95 \%$ confidence interval was $[0.76,0.97]$ and the estimated probability of success is 0.9 . For the 24 successes (third quartile), the $95 \%$ confidence interval was [0.43, 0.75] and the estimated probability of success was 0.6 . In all cases the estimated probability of success fell around the mid point of the interval, which is a good result; moreover, the intervals are narrow and their first endpoint was higher than 0.5 in the first two cases, which means that we are $95 \%$ sure that the success happens in more than $50 \%$ of the cases. In other words, the success is more likely than the failure, which means a good result for the evaluation.

Random filler vs. remaining 39 random fillers As explained on page 112, we compared the probability assigned by the system to each verb's random filler with the mean of the probabilities assigned to the 39 random fillers extracted for the other verbs. In 24 cases out of 40 the former probability was higher than the latter probability. A binomial test on these data shows that the proportion of successes was significantly higher than the proportion of failures: confidence interval [ $0.433,0.751$ ], estimated probability of success 0.6.

## B.9.1 Corpus frequencies

Maximum-frequency fillers vs. minimum-frequency fillers We extracted the fillers of the 40 verb-slot pairs from two automatically parsed corpora, as described in section 5.4.2 in section 5.4.2.1. We then calculated the SP probabilities of the maximum-frequency fillers (which were one for each verb-slot pair) and the mean of the SP probabilities of the minimum-frequency fillers, as shown in table B.13. The cases when the first was higher than the second were considered as successes. For example, the fillers for the slot 'A_(in)Obj[acc]' of intro 'enter' are: mundus 'world' (22), homo 'man' (3), gloria 'glory' (3), paradisus 'heaven' (3), orbis 'world' (2), definitio 'boundary' (1), cognitio 'knowledge' (1), terra 'earth' (1), constitutio 'regulation' (1), possessio 'possession' (1). The maximum-frequency filler is mundus, which is assigned the SP probability of 0.0013 . The minimum-frequency fillers are definitio, cognitio, terra, constitutio, and possessio; the mean of their SP probabilities is 0.0011 . Intro is thus an instance of success.

Of the 40 pairs, 27 were successes. Among the failures we considered ascendo for which only one filler was extracted. A binomial test on these data gave the $95 \%$ confidence interval $[0.509,0.814]$ and an estimated probability of success of 0.675 .

Correlation test We compared the frequency distributions found in the automatically parsed corpora with the SP probability distribution given by our model. For this,

| verb | slot | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| accido2 | A_Obj[dat] | 7 | 0.00048 | 1 | 0.00047 | 1 | 0.00044 | 0.00029 | 1 |
| agito | A_Obj[acc] | 7 | 0.00072 | 1 | 0.00036 | 1 | 0.00038 | 0.00029 | 1 |
| ascendo | A_(de)Obj[abl] | 1 | 0.00026 | 1 | 0.00026 | 0 | 0.00026 | 0.00029 | 0 |
| canto | A_Sb | 5 | 0.00013 | 2 | 0.0014 | 0 | 0.00054 | 0.00024 | 1 |
| cognosco | A_Sb | 84 | 0.0010 | 1 | 0.00046 | 1 | 0.00057 | 0.00026 | 1 |
| comparo | P_(ad)Obj[acc] | 13 | 0.00066 | 1 | 0.00061 | 1 | 0.00064 | 0.00028 | 1 |
| constituo | A_Obj[acc] | 34 | 0.0034 | 1 | 0.00041 | 1 | 0.00049 | 0.00028 | 1 |
| convenio | A_Obj[dat] | 64 | 0.00064 | 1 | 0.00046 | 1 | 0.00050 | 0.00027 | 1 |
| cupio | A_Obj[acc] | 4 | 0.00070 | 1 | 0.00051 | 1 | 0.00049 | 0.00026 | 1 |
| determino | A_Sb | 96 | $2.8361 \mathrm{e}-09$ | 1 | 0.00029 | 0 | 0.0003 | 0.00034 | 0 |
| dico2 | A_Obj[acc] | 81 | 0.00196 | 1 | 0.00031 | 1 | 0.00038 | 0.00027 | 1 |
| dimitto | A_Sb | 4 | 0.00126 | 1 | 0.00078 | 1 | 0.00080 | 0.00024 | 1 |
| do | A_Obj[dat] | 30 | 0.00139 | 1 | 0.00040 | 1 | 0.00041 | 0.00028 | 1 |
| facio | A_Obj[acc] | 55 | 5.6077e-05 | 1 | 0.00030 | 0 | 0.00035 | 0.00027 | 1 |
| fero | A_Obj[acc] | 47 | 0.00079 | 1 | 0.00034 | 1 | 0.00038 | 0.00028 | 1 |
| induco | A_(in) $\mathrm{Obj}[\mathrm{acc}]$ | 3 | 0.00090 | 1 | 0.00066 | 1 | 0.00064 | 0.00029 | 1 |
| intellego | A_Sb | 150 | 0.00194 | 1 | 0.00032 | 1 | 0.00025 | 0.00035 | 0 |
| intro | A_(in)Obj[acc] | 22 | 0.00127 | 1 | 0.00107 | 1 | 0.00099 | 0.00029 | 1 |
| invenio | A_Obj[acc] | 13 | 0.00054 | 1 | 0.00035 | 1 | 0.00039 | 0.00027 | 1 |
| maneo | A_Sb | 25 | 0.00040 | 1 | 0.00040 | 0 | 0.00051 | 0.00028 | 1 |
| moveo | A_(ad)Obj[acc] | 6 | 0.00105 | 1 | 0.00035 | 1 | 0.00050 | 0.00027 | 1 |
| multiplico | A_Obj[acc] | 7 | $1.556 \mathrm{e}-12$ | 1 | 0.00043 | 0 | 0.00035 | 0.00029 | 1 |
| nego | A_Obj[acc] | 7 | 0.00043 | 1 | 0.00035 | 1 | 0.00042 | 0.00030 | 1 |
| paro | A_Obj[acc] | 23 | 0.00030 | 1 | 0.00037 | 0 | 0.00038 | 0.00030 | 1 |
| pateo | A_Sb | 272 | $7.8697 \mathrm{e}-05$ | 1 | 0.00058 | 0 | 0.00058 | 0.00029 | 1 |
| pono | A_Obj[acc] | 48 | 0.00101 | 1 | 0.00035 | 1 | 0.00042 | 0.00026 | 1 |
| praecedo | A_Obj[acc] | 23 | 0.00059 | 1 | 0.00039 | 1 | 0.00046 | 0.00028 | 1 |
| praedico | A_Sb | 8 | 0.00040 | 1 | 0.00037 | 1 | 0.00047 | 0.00026 | 1 |
| profero | A_Sb | 20 | $2.0889 \mathrm{e}-10$ | 1 | 0.00078 | 0 | 0.00072 | 0.00024 | 1 |
| recipio | A_Obj[acc] | 29 | 0.00153 | 1 | 0.00043 | 1 | 0.00050 | 0.00028 | 1 |
| refero | A_(ad)Obj[acc] | 7 | 1.1266e-08 | 1 | 0.00045 | 0 | 0.00047 | 0.00034 | 1 |
| resulto | A_Sb | 7 | $8.547 \mathrm{e}-05$ | 1 | 0.00037 | 0 | 0.00056 | 0.00028 | 1 |
| scio | A_Sb | 16 | 0.00089 | 1 | 0.00044 | 1 | 0.00047 | 0.00021 | 1 |
| sequor | A_Obj[acc] | 5 | 0.00111 | 1 | 0.00044 | 1 | 0.00048 | 0.00028 | 1 |
| subeo | A_Obj[acc] | 7 | 0.00074 | 1 | 0.00035 | 1 | 0.00039 | 0.00029 | 1 |
| sum | A_Sb | 638 | 0.00122 | 1 | 0.00024 | 1 | 0.00027 | 0.00032 | 0 |
| tento | A_Obj[acc] | 7 | 0.00174 | 1 | 0.00049 | 1 | 0.00054 | 0.00028 | 1 |
| timeo | A_Sb | 10 | 0.00017 | 1 | 0.00043 | 0 | 0.00056 | 0.00026 | 1 |
| venio | A_(ad)Obj[acc] | 13 | 0.00027 | 1 | 0.00041 | 0 | 0.00046 | 0.00029 | 1 |
| volo | A_Sb | 78 | 0.00077 | 1 | 0.00047 | 1 | 0.00051 | 0.00026 | 1 |

Table B.13: Comparison between the frequencies and the SP probabilities of the maximum-frequency and minimum-frequency fillers extracted from the parsed corpora. The fourth column contains the maximum frequency of the fillers; the fifth column contains the probability of the maximum-frequency filler; the sixth column contains the mean of the minimum-frequency fillers, the seventh column contains the mean of the probabilities of the minimum-frequency fillers; the eighth column contains the binary answers to the question "is the probability of the maximum-frequency filler higher than the mean of the probabilities of the minimum-frequency fillers?"; the ninth column contains the mean of the probabilities assigned to the fillers; the tenth column contains the mean of the probabilities assigned to the "non fillers"; the eleventh column to the question "is the mean of the probabilities of the fillers higher than the mean of the probabilities of the "non fillers"?"
we performed a Spearman rank correlation test computed in R (R Development Core Team, 2008). This test converts the frequencies to ranks and is an ordinal version of the Pearson correlation test (Hinton, 2004, pp. 279-281).

The Spearman test calculates the Spearman coefficient $r$ and then decides on its significance based on tables of critical values. The formula for $r$ is

$$
r=1-\frac{6 \sum_{i=1}^{N} D_{i}^{2}}{N^{3}-N}
$$

where $\left(D_{1}, \ldots, D_{N}\right)$ is the vector containing the differences between the two ranks and $N$ is the number of observations.

For example, table B. 14 shows the two distributions for the arguments of intro 'enter' for the slot 'A_(in) $\mathrm{Obj}[\mathrm{acc}]$ '. The SP probabilities displayed correspond to the fillers for intro in the parsed corpora. The two last columns contain the ranks of the two distributions. Note that two similar nouns like mundus 'world' and terra 'earth' do not appear in the same position in the two rankings: mundus is first by frequency in the parsed corpora and terra is first in the SP probability distribution. This may be due to the their semantic score: $\frac{15}{90}=0.17$ (see sections B.7.2.2 and B.7.2.3). This score is low because the WN-based measure of similarity we adopted does not distinguish between common and less common senses of words.

| Index | filler | frequency | SP prob. | freq. rank | proba. rank |
| ---: | :--- | :--- | :--- | ---: | ---: |
| 1 | mundus | 22 | $1.27 \mathrm{e}-03$ | 10 | 7 |
| 2 | homo | 3 | $3.35 \mathrm{e}-04$ | 9 | 5 |
| 3 | gloria | 3 | $1.39 \mathrm{e}-03$ | 8 | 8 |
| 4 | paradisus | 3 | $1.69 \mathrm{e}-10$ | 7 | 1 |
| 5 | orbis | 2 | $1.57 \mathrm{e}-03$ | 6 | 9 |
| 6 | definitio | 1 | $3.93 \mathrm{e}-10$ | 5 | 2 |
| 7 | cognitio | 1 | $2.60 \mathrm{e}-04$ | 4 | 4 |
| 8 | terra | 1 | $4.76 \mathrm{e}-03$ | 3 | 10 |
| 9 | constitutio | 1 | $3.44 \mathrm{e}-04$ | 2 | 6 |
| 10 | possessio | 1 | $4.98 \mathrm{e}-10$ | 1 | 3 |

Table B.14: Corpus frequency and SP probabilities for the 'A_(in)Obj[acc]'-arguments of intro 'enter'.

In this case, $N=10, D=(10-7,9-5,8-8,7-1,6-9,5-2,4-4,3-10,2-6,1-3)=$ $(3,4,0,6,-3,3,0,-7,-4,-2), D^{2}=\left(3^{2}, 4^{2}, 0^{2}, 6^{2},(-3)^{2}, 3^{2}, 0^{2},(-7)^{2},(-4)^{2},(-2)^{2}\right)=$ $(9,16,0,36,9,9,0,49,16,4)$ and $\sum D^{2}=148$. Therefore, $r=1-\frac{6 \cdot 148}{10^{3}-10}=0.10$. At this point, we looked up the critical value for $N=10$ and 0.01 level of significance in the table of critical values presented in Hinton (2004, p. 373): we found a critical value of 0.746 , which is higher than the $r$ we calculated. This means that there is no significant correlation in the two ranking given by the two distributions.

Table B. 15 shows the correlation coefficients, the $p$-values and the statistics of the Spearman tests run on the 40 verb-slot pairs. For 12 pairs (in bold in table B.15) we got $p$-values lower or equal to 0.05 and $r$ coefficients between 0.13 (fero) and 0.43 (moveo). For the remaining 28 pairs we got a $p$-value higher than 0.05 , which implies that in 28 out of 40 cases the test gave a non-significant correlation at the 0.05 level. If the had chosen the level of 0.01 , the number of significant cases would have been 4 out of 40 . A binomial test on 12 successes gives a $95 \%$ confidence interval of $[0.166,0.465]$ and an
estimated probability of success of 0.3 . Since the interval is narrow and its lower margin is lower than 0.5 , we conclude that we are $95 \%$ confident that the 12 successes occurred by chance. ${ }^{20}$

Fillers vs. "non-fillers" As a last measure of goodness for our system, we compared the mean of $P_{a}(f \mid v)$ ( $f$ is a filler for $(v, a)$ in the parsed corpora) with the mean of $P_{a}(n \mid v)$ ( $n$ is a filler for other pairs than $(v, n)$ ). The result gave 36 successes, where the mean of $P_{a}(f \mid v)$ was higher than the mean of $P_{a}(n \mid v)$. A binomial test on 36 successes out of 40 provides a $95 \%$ confidence interval of [ $0.763,0.972$ ] and an estimated probability of success of 0.9 , which says that the proportion of successes is significantly higher than the proporion of failures.

## B.9.2 Distributional similarity

We performed an evaluation of the "distributional" version of the algorithm described in section 5.4.3.

Central tendency First, we compared the probability assigned by this algorithm to the random fillers with the mean, median and value at the third quartile of the distribution for each verb-slot pair.

In 18 of the 39 cases the probability was higher than the mean, showing that the proportion of successes was not higher than the proportion of failures, as confirmed by a binomial test: $95 \%$ confidence interval $=[0.301,0.628]$, estimated probability of success $=0.461$. In 35 cases the probability was higher than the median, which says that the proportion of successes is significantly higher than the proportion of failures: $95 \%$ confidence interval $=[0.758,0.971]$, estimated probability of success $=0.897$. Finally, in 17 cases the probability was higher than the value of the third quartile, which means a significantly lower probability of success: $95 \%$ confidence interval $=[0.278,0.604]$, estimated probability of success $=0.436$.

Fillers from parsed corpora We extracted the fillers of the 40 verb-slot pairs from two automatically parsed corpora, and then calculated the SP probabilities of the maximum-frequency fillers and the mean of the SP probabilities of the minimumfrequency fillers. The cases when the first was higher than the second were considered as successes.

Of the 39 pairs, 18 were successes. Among the failures we considered ascendo for which only one filler was extracted. A binomial test on these data gave the $95 \%$ confidence interval $[0.301,0.628]$ and an estimated probability of success of 0.462 .

For each pair $(v, a)$ we compared the mean of the probabilities of the fillers with the mean of the probabilities of the fillers for the other pairs and considered as successes those cases where the former was higher than the latter. We obtained 32 successes and a binomial test of these data gave a $95 \%$ confidence interval of $[0.665,0.925]$ and an estimated probability of success of 0.82 .

Correlation test We used a Spearman correlation test to compare the frequency distributions found in the automatically parsed corpora with the SP probability distribution given by our model.

[^81]Table B. 16 shows the correlation coefficients, the $p$-values and the statistics of Spearman tests run on the 40 verb-slot pairs. For 16 pairs (in bold in table B.15) we got $p$-values lower or equal to 0.05 and $r$ coefficients between 0.04 (sum) and 1 (canto). ${ }^{21}$ For the remaining 24 pairs we got $p$-values higher than 0.05 , which implies a non-significant correlation at the 0.05 level. A binomial test on 16 successes gives a $95 \%$ confidence interval of $[0.249,0.567]$, and an estimated probability of success of 0.4 , which means that we are $95 \%$ confident that the 12 successes occurred by chance.

[^82]| verb | $r$ | p -value | statistic |
| ---: | :--- | :--- | :--- |
| accido2 | -0.078 | 0.615 | 15295.941 |
| agito | 0.117 | 0.443 | 13400.823 |
| canto | -0.5 | 0.667 | 6 |
| cognosco | 0.230 | 0.0172 | 157228.776 |
| comparo | 0.239 | 0.062 | 30231.272 |
| constituo | 0.247 | 0.004 | 308573.364 |
| convenio | 0.095 | 0.284 | 323784.481 |
| cupio | -0.106 | 0.429 | 35951.592 |
| determino | -0.067 | 0.644 | 22218.496 |
| dico2 | 0.235 | $2.115 \mathrm{e}-07$ | 13840039.950 |
| dimitto | 0.112 | 0.679 | 603.790 |
| do | 0.165 | 0.019 | 1181899.648 |
| facio | 0.201 | $1.560 \mathrm{e}-07$ | 39867702.500 |
| fero | 0.134 | 0.011 | 6731259.403 |
| induco | -0.087 | 0.693 | 2199.923 |
| intellego | -0.002 | 0.989 | 54837.015 |
| intro | 0.111 | 0.759 | 146.610 |
| invenio | 0.181 | 0.012 | 965591.911 |
| maneo | 0.307 | 0.0007 | 194536.3307 |
| moveo | 0.430 | 0.041 | 1153.452 |
| multiplico | -0.456 | 0.217 | 174.772 |
| nego | 0.155 | 0.105 | 187329.390 |
| paro | 0.0128 | 0.906 | 115985.660 |
| pateo | 0.059 | 0.677 | 23351.174 |
| pono | 0.203 | $7.415 \mathrm{e}-05$ | 7120781.734 |
| praecedo | 0.260 | 0.004 | 235252.190 |
| praedico | 0.414 | 0.044 | 1347.408 |
| profero | -0.185 | 0.609 | 195.536 |
| recipio | 0.197 | 0.004 | 1310847.672 |
| refero | 0.262 | 0.015 | 78173.180 |
| resulto | 0.055 | 0.866 | 270.388 |
| scio | 0.071 | 0.584 | 36894.184 |
| sequor | 0.287 | 0.062 | 9447.610 |
| subeo | 0.073 | 0.475 | 140924.768 |
| sum | 0.135 | $4.033 \mathrm{e}-05$ | 110032847.665 |
| tento | 0.205 | 0.102 | 36393.017 |
| timeo | 0.191 | 0.513 | 368.151 |
| venio | 0.160 | 0.050 | 482171.470 |
| volo | 0.222 | 0.039 | 85351.370 |
|  |  |  |  |

Table B.15: Correlation coefficients $r, p$-values and statistics of the Spearman correlation test on the corpus frequency distribution and the SP probability distribution given by our model.

| verb | $r$ | $p$-value | statistic |
| ---: | :--- | :--- | :--- |
| accido2 | 0.209 | 0.0872 | $4.14 \mathrm{e}+04$ |
| agito | 0.135 | 0.233 | $7.38 \mathrm{e}+04$ |
| canto | 1 | $\mathbf{0}$ | 0 |
| cognosco | 0.229 | $\mathbf{0 . 0 0 0 6 5 3}$ | $1.33 \mathrm{e}+06$ |
| comparo | 0.492 | $\mathbf{1 . 2 9 e - 0 6}$ | $5.57 \mathrm{e}+04$ |
| constituo | 0.208 | $\mathbf{0 . 0 0 3 4}$ | $1.01 \mathrm{e}+06$ |
| convenio | 0.402 | $\mathbf{1 . 8 5 e - 0 9}$ | $8.98 \mathrm{e}+05$ |
| cupio | 0.074 | 0.49 | $1.09 \mathrm{e}+05$ |
| determino | 0.225 | $\mathbf{0 . 0 3 8 7}$ | $7.93 \mathrm{e}+04$ |
| dico2 | 0.119 | $\mathbf{0 . 0 0 0 2 0 7}$ | $1.3 \mathrm{e}+08$ |
| dimitto | 0.074 | 0.684 | $5.54 \mathrm{e}+03$ |
| do | 0.150 | $\mathbf{0 . 0 0 1 4 4}$ | $1.27 \mathrm{e}+07$ |
| facio | 0.181 | $\mathbf{1 . 4 3 e - 1 0}$ | $2.56 \mathrm{e}+08$ |
| fero | 0.101 | $\mathbf{0 . 0 1 7 4}$ | $2.59 \mathrm{e}+07$ |
| induco | 0.081 | 0.683 | $3.36 \mathrm{e}+03$ |
| intellego | 0.316 | $\mathbf{0 . 0 0 0 1 0 2}$ | $3.55 \mathrm{e}+05$ |
| intro | 0.355 | 0.284 | 142 |
| invenio | 0.110 | 0.0601 | $3.69 \mathrm{e}+06$ |
| maneo | 0.370 | $\mathbf{5 . 2 9 e - 0 8}$ | $8.92 \mathrm{e}+05$ |
| moveo | 0.103 | 0.586 | $4.03 \mathrm{e}+03$ |
| multiplico | -0.149 | 0.663 | 253 |
| nego | 0.130 | 0.104 | $5.72 \mathrm{e}+05$ |
| paro | 0.094 | 0.279 | $3.63 \mathrm{e}+05$ |
| pateo | 0.089 | 0.388 | $1.39 \mathrm{e}+05$ |
| pono | 0.256 | $\mathbf{3 . 6 6 e - 1 0}$ | $2.44 \mathrm{e}+07$ |
| praecedo | 0.070 | 0.366 | $7.22 \mathrm{e}+05$ |
| praedico | 0.185 | 0.199 | $1.7 \mathrm{e}+04$ |
| profero | 0.560 | $\mathbf{0 . 0 2 4 1}$ | 299 |
| recipio | 0.050 | 0.381 | $4.95 \mathrm{e}+06$ |
| refero | 0.048 | 0.614 | $2.35 \mathrm{e}+05$ |
| resulto | 0.226 | 0.338 | $1.03 \mathrm{e}+03$ |
| scio | 0.126 | 0.203 | $1.64 \mathrm{e}+05$ |
| sequor | -0.057 | 0.664 | $4 \mathrm{e}+04$ |
| subeo | 0.108 | 0.224 | $3.12 \mathrm{e}+05$ |
| sum | 0.050 | $\mathbf{0 . 0 1 1 2}$ | $2.74 \mathrm{e}+09$ |
| tento | -0.083 | 0.43 | $1.41 \mathrm{e}+05$ |
| timeo | 0.290 | 0.114 | $3.52 \mathrm{e}+03$ |
| venio | 0.242 | $\mathbf{0 . 0 0 0 1 0 4}$ | $2.02 \mathrm{e}+06$ |
| volo | 0.081 | 0.285 | $8.35 \mathrm{e}+05$ |
|  |  |  |  |

Table B.16: Correlation coefficients $r, p$-values and statistics of the Spearman correlation test on the corpus frequency distribution and the SP probability distribution given by the distributional version of our model.

## Appendix C

## Statistical background to chapter 7

## C. 1 Introduction

This appendix illustrates the statistical techniques employed in chapter 7 and gives details of our corpus analyses. First, we present the $\chi^{2}$ test (section C.2), then linear models (section C.3), and finally, Correspondence Analysis (section C.4).

## C. 2 Pearson's $\chi^{2}$ test of independence

In this section we will briefly describe Pearson's $\chi^{2}$ test of independence and discuss its use; we used this test in the first study on Latin preverbs reported on in section 7.4. For a more in-depth description and mathematical details on this test, see Hinton (2004, pp. 248-259).

Pearson's $\chi^{2}$ test belongs to the set of techniques used in hypothesis testing. As mentioned on page 140, hypothesis testing aims at testing a formal hypothesis by attempting to reject it. Research questions often involve decisions about populations and their distributions, for example the frequencies of word order patterns in classical and late Latin. The populations are in most cases not available for direct study; however, with hypothesis testing it is possible to compare samples from the populations, and decide whether or not the populations differ based on these samples.

Hypothesis testing is based on setting up a null hypothesis $\left(H_{0}\right)$, and comparing it with an alternative hypothesis $\left(H_{1}\right) . H_{0}$ can be formulated in a number of ways; however, it always specifies the most conservative scenario such as "no difference" or "no effect". In the word order example above, we can formulate the null hypothesis that the two samples (frequencies of word order patterns from classical and late Latin, respectively) are drawn from the same population. The alternative hypothesis would be that they are drawn from different populations. To model the populations, a mathematical distribution with known properties, such as the $\chi^{2}$ distribution, can be used (see details below). It can be shown that for frequencies of categories (i.e. nominal data) the $\chi^{2}$ distribution is a fair model. Thus, we have a formal device for comparing the two hypotheses with the observed frequencies in the samples.

The analysis proceeds by a priori deciding on a significance level (commonly 0.05

| author | construction A | construction B | row totals |
| :--- | :---: | :---: | ---: |
| Plautus | 15 | 24 | 39 |
| Thomas | 2 | 53 | 55 |
| column totals | 17 | 77 | 94 |

Table C.1: Frequency counts of two constructions in Plautus and Thomas in our corpora.
or $5 \%$ in the humanities and social sciences), which is the threshold at which we reject $H_{0}$; then we either reject or do not reject the null hypothesis. Put differently, we only reject the null hypothesis when the data seem extremely unlikely in the light of the assumptions of $H_{0}$, so that the probability of taking the right decision is at least $95 \%$.

Pearson's $\chi^{2}$ test works on frequency data in the form of a table, where two variables are assigned to rows and columns, respectively. We will illustrate the case of two-bytwo tables, where two variables are divided into two categories; however, the discussion applies to a higher number of categories as well.

As an example, table C. 1 gives the frequencies of two syntactic constructions (A and B) in two authors (Plautus and Thomas), taken from the dataset of our first study (see section 7.4). In this case, the two variables we categorize our data in terms of are 'construction' and 'author'.

A Pearson $\chi^{2}$ test for independence formally tests the null hypothesis that rows and columns are independent of each other, i. e. all the observations in the table follow a random uniform discrete distribution: the chi-squared distribution. In the current example, the null hypothesis states that the frequencies of constructions A and B do not depend on the particular author they occur in, i. e. they are drawn from the same $\chi^{2}$-distributed population. ${ }^{1}$ The total number of observations is 39 for Plautus and 55 for Thomas, so even under the null hypothesis we do not expect the same observed frequencies for the two constructions in the two authors; rather, we would expect the same proportion of each construction category in each author. For example, the 24 occurrences of construction A in Plautus correspond to $\frac{24}{15+24}=0.62$ or $62 \%$ of the total frequency in Plautus. If there was no difference between the two authors in terms of the constructions used, then the 17 total occurrences of construction A should be distributed between Plautus and Thomas in proportion to the authors' relative number: the proportion of Plautus is $\frac{39}{94}=0.415$, so we would expect $0.41 \cdot 17=7.06$ instances of construction A in Plautus.

The expected value of each cell $(i, j)(i=1, \ldots, n r=$ number of rows and $j=$ $1, \ldots, n c=$ number of columns) in the table is calculated as

$$
\text { expected }_{i, j}=\frac{\text { row } i \text { total } \cdot \text { column } j \text { total }}{\text { overall total }}
$$

Table C. 2 contains the expected counts for each cell in table C.1.
The assumption of a Pearson's $\chi^{2}$ test for independence is that the observations are independent and randomly extracted from the population, and that their expected frequencies are higher than 5 (Hinton, 2004, p. 258), as in our example.

Under the null hypothesis, the observed values are very close to the expected values, although they may not coincide with them due to random variation. The deviation from the expected behaviour is measured by the $\chi^{2}$ statistic:

[^83]| author | construction A | construction B |
| :--- | :---: | :---: |
| Plautus | 7.06 | 31.94 |
| Thomas | 9.95 | 45.05 |

Table C.2: Expected frequency counts from table C.1.

$$
\chi^{2}=\sum_{i=1, \ldots, n r ; j=1, \ldots, n c} \frac{(O-E)^{2}}{E}
$$

where $O$ refers to the observed value in a cell and $E$ to its expected value. In the previous example, $\chi^{2}=\frac{(15-7.05)^{2}}{7.05}+\frac{(24-31.94)^{2}}{31.94}+\frac{(2-9.95)^{2}}{9.95}+\frac{(53-45.05)^{2}}{45.05}=18.69$.

A decision on the statistical significance of this value is taken by comparing it with its table of critical values. ${ }^{2}$ For example, given the $\chi^{2}$ statistic just calculated and the number of rows and columns in our input table, we find a critical value of 3.84: since the calculated value is greater than this critical value, we can reject the null hypothesis with a probability of $5.117 \mathrm{e}-05$. This is the $p$-value given by the test; if it is lower than the predefined level of significance (0.05), than we can be sure that in more than $95 \%$ of the cases the null hypothesis is wrong. Therefore, the two authors use the two constructions in ways which are significantly different at the 0.05 level.

In addition to the $p$-value, which states whether the result is significant or not, i.e. if the two variables are significantly independent or not, it is relevant to consider the effect size, which accounts for the size of the detected independence. The effect size is an important indicator about the size of our sample: in fact, high frequencies - as well as a large number of categories, i. e. columns in the input table- tend to magnify small deviations, as pointed out in Mosteller (1968, p. 2) about the Brown Corpus. The effect size factors out the sample size, and provides a standardized measure of how large the difference is.

A measure of effect size for $2 \times 2$ tables is $\phi$ :

$$
\phi=\sqrt{\frac{\chi^{2}}{n}}
$$

where $n$ is the total sum of the cell frequencies and $\phi$ varies between 0 and 1 . In this example, $\phi=0.42$.

For larger tables, Cramér $V$ is used, which approximately varies between 0 and 1 :

$$
V=\sqrt{\frac{\chi^{2}}{n \cdot(\min \{n r, n c\}-1)}}
$$

According to Cohen (1988, pp. 222-227) values of $\phi$ (or $V$ ) around 0.10 are to be considered a small effect, values around 0.30 a medium effect and values around and above 0.50 a large effect. In our example, we hence get a medium-to-high effect.

[^84]
## C. 3 Linear regression models

This section provides an overview of linear regression models, which were employed in the second study on Latin preverbs reported on in section 7.4.3. We refer to Gorard (2003, pp. 202-226) and Hinton (2004, pp. 262-281, 284-294) for a more detailed discussion on the possible problems arising with these models.

Section C.3.1 introduces linear functions, while section C.3.2 explains the main ideas behing linear regression. Sections C.3.3 and C.3.4 go into generalized linear models and mixed effect models, respectively.

## C.3.1 Linear functions

Linear regression models approximate the relation between the variables in the data with a linear function. In this section we will define linear functions and give a geometrical interpretation of them.

Let us consider a two-dimensional space, identified by the $x$ and $y$ axes. Let us also suppose that our data are composed of a set of points, each of which is uniquely associated with a pair of coordinates $(x, y)$. We will focus on the particular case when the points lie along a line: see figure C.1.


Figure C.1: Example of the line with equation $y=2 x+1$ and intercept $b=1$. The point $P=(1.5,4)$ is represented.

A line can be defined by a linear function, which gives the rule for obtaining a corresponding $y$ from each $x$ : this function has a simple expression:

$$
\begin{equation*}
y=a x+b \tag{C.1}
\end{equation*}
$$

$b$ is called intercept and corresponds to the ordinate of the intersection point between the line and the $y$ axis, while $a$ is called slope and measures how steep the line is. For the line in figure C. $1, b=1$ and $a=2$. Note that, for each $x, y$ can be calculated by applying formula (C.1): for example, if $x=1.5$, then $y=2 \cdot 1.5+1=4$ (see point $P$ in figure C.1).

## C.3.2 Linear regression

This section gives an overview on simple (section C.3.2.1) and multiple (section C.3.2.2) linear regression.

In data analysis, regression models aim at finding the best linear function approximating the data at hand. This way, it is possible to predict one variable (called response) as a linear function of the other(s) (called predictors). Furthermore, the scores calculated by applying the model to the data should be reasonably similar to the scores actually observed in the data, so that is the error is minimised. In other words, in a good model the predictors explain most of the variability in the response.

The assumptions of a regression analysis are numerous (see, for example, Gorard (2003, p. 213)) and they require, among other things, that the data come from a random sample, the response is approximately linearly related to the predictors, that the differences between calculated and observed scores are approximately normally distributed, ${ }^{3}$ and that the predictors are uncorrelated.

Finding a regression model for our data is equivalent to finding a correlation between the variables. For example, we may find that the average length of the utterances spoken by babies and the age in months are correlated (or linearly related), since the length of the utterances increases by progressing in time (positive slope).

It should pointed out, though, that correlation is distinguished from causality. As Gorard (2003, p. 205) notes, finding that children with bigger feet also tend to spell better does not explain why this is the case and may be due to a third variable affecting the two previous ones, such as age.

## C.3.2.1 Simple linear regression

Let us suppose that our data consist of pairs of scores, i. e. vectors: $X$ and $Y$. First, we need to set which variable to predict - the response, say $Y$ - and which variable to use as predictor, say $X$. We then aim at finding the linear function $L$ such that the $Y$ scores can be obtained by applying $L$ to the $X$ scores, except for some random variation.

$$
Y=L(X)+E
$$

or

$$
Y=Y^{\prime}+E
$$

where $Y^{\prime}$ is a straight line and $E$ is the error, which causes the data points to deviate from the line. $Y^{\prime}=a X+b$ is the regression line: the best regression line is considered to be the line that gives the best fit to the data, i.e. the line that minimises the error $E$. What a regression analysis does, then, is to test if the corresponding linear model is a good model for the data at hand.

[^85]| observation | speed | distance |
| ---: | :--- | :--- |
| $P_{1}$ | 4 | 2 |
| $P_{2}$ | 7 | 4 |
| $P_{3}$ | 9 | 10 |
| $P_{4}$ | 11 | 17 |
| $P_{5}$ | 12 | 20 |
| $P_{6}$ | 13 | 26 |

Table C.3: Example containing values for car speed and distance to stop.
$E=Y-Y^{\prime}$ can be seen as the distance between the data points and the regression line and collects the so-called residual values. Along this reasoning, $E$ is found through the procedure of least squares, which gives the constants $a$ and $b$ defining $Y^{\prime}$ :

$$
a=\frac{S P}{S S_{X}}
$$

and

$$
b=\bar{Y}-B \bar{X}
$$

where $\bar{X}$ and $\bar{Y}$ are the means of the values of $X$ and $Y, S P=\sum(X-\bar{X})(Y-\bar{Y})$ and $S S_{X}=\sum(X-\bar{X})^{2}$.

For example, consider the points in figure C.2, a subset of the 'cars' dataset contained in the $R$ manual ( R Development Core Team, 2008) and representing the speed of cars on the $x$ axis and the corresponding distance to stop on the $y$ axis. The points' coordinates are displayed in table C.3. As shown in figure C.2, the points are scattered in the space, but a linear pattern can be detected. If we apply the least square method to this example, we get $\bar{X}=9.33, \bar{Y}=13.17, S P=153.67$, and $S S_{X}=53.33$, thus $a=2.68$ and $b=-11.85$. The regression line is then defined by $Y^{\prime}=2.68 X-11.85$ and represented in figure C.3.

## C.3.2.2 Multiple linear regression

Multiple linear regression is the generalization of simple linear regression to the case when more than two variables are involved. As in simple regression, one variable must be chosen as the response $Y$ and will be predicted by other $k$ variables, the predictors $X_{1}, \ldots, X_{k}$. Then, we aim at solving the linear equation:

$$
\begin{equation*}
Y=a+b_{1} X_{1}+b_{2} X_{2}+\ldots+b_{k} X_{k} \tag{C.2}
\end{equation*}
$$

If $n$ is the number of observations, both the response and the predictors are $n$ dimensional vectors: $Y=\left(y_{1}, y_{2}, \ldots, y_{n}\right), X_{1}=\left(x_{1,1}, x_{1,2}, \ldots, x_{1, n}\right), \ldots$, until $X_{k}=$ ( $x_{k, 1}, x_{k, 2}, \ldots, x_{k, n}$ ). Equation (C.2) translates into a system of $n$ equations:

$$
y_{i}=a+b_{1} x_{1, i}+b_{2} x_{2, i}+\ldots+b_{k} x_{k, i}
$$

for $i=1, \ldots, n$.
For example, we may be interested in predicting the salary of people based on two predictors: years of education and age; that is, we want to express the salary $(Y)$ in our dataset $(D)$ as linearly depending on the length of the education $\left(X_{1}\right)$ and the age $\left(X_{2}\right)$ of a person. In this example, the linear equation to be solved is:

$$
\begin{equation*}
Y=a+b_{1} X_{1}+b_{2} X_{2} \tag{C.3}
\end{equation*}
$$



Figure C.2: Data points from table C. 3 relative to speed of cars and their distance to stop.


Figure C.3: Regression line of the data points plotted in figure C.2.
or

$$
\begin{equation*}
\text { salary }=a+b_{\text {education }} \cdot \text { education }+b_{\text {age }} \cdot \text { age } \tag{C.4}
\end{equation*}
$$

Explaining the way equations (C.2) and (C.3) are solved falls outside the scope of this dissertation. Instead, we will illustrate the way this problem is formulated in R and the kind of generated output, since this is how we used multiple regression models in the second study on preverbs presented in section 7.4.3, and whose results are reported in section C.3.5.

The function $\operatorname{lm}()$ in R fits a linear model to a dataset in the form response $\sim$ predictors, where the predictors are listed with the ' + ' sign; thus the model represented by equation (C.4) is calculated with

```
lm(construction ~ author + verb, data = D)
```

R produces an output summary containing the distribution of residuals (which are assumed to follow a random distribution), a coefficient for the intercept (a) and a coefficient for the slope $\left(b_{i}\right)$ of each predictor. These coefficients are followed by numbers attesting the statistical significance and the reliability of the estimated coefficients. In addition to inspecting the model's summary, an appropriate model diagnostic should be carried out that checks its fit to the data by plotting its residuals. We will examine the output summary and the diagnostic plots more in detail case by case in section C.3.5, after outlining the main subtypes of models we used in sections C.3.3 and C.3.4.

## C.3.3 Generalized Linear Models

A Generalized Linear Model (GLM) is a generalization of a linear regression model to deal with a wider range of cases. A GLM is specified by a family distribution for the residuals. A particular class of GLMs are logistic models, which are used when the response is a binary variable (i.e. it takes only two values, 0 and 1 , failure and success, for example a construction with or without a preposition). In addition, a logistic model requires that the response has a binomial error distribution. ${ }^{4}$

In logistic models our goal is to estimate the probability of one of the two possible outcomes (e.g. absence or presence of preposition), given the predictors (e.g. author and verb). Instead of being modelled directly, the probabilities are modelled through a so-called link function, which is the logit transformation of the probability:

$$
\operatorname{logit}(x)=\ln \left(\frac{x}{1-x}\right)
$$

The logit function maps the range $[0,1]$ to the range $[-\infty,+\infty]$. Once the logit coefficients are produced by the model, it is necessary to back transform them with the inverse of the logit function to interpret them as probabilities.

The R syntax for a logistic GLM modelling the binary variable 'preposition' (taking two values) on the predictors 'author' and 'verb' is:

$$
\text { glm(preposition } \sim \text { author }+ \text { verb, data }=\text { D, family = binomial) }
$$

## C.3.4 Mixed effect models

Mixed effect models are an expansion of classical linear regression involving fixed effects (i. e. predictors) and random effects. The combination of fixed and random effects allows us to account for both individual and group level variation in a model. This turns out particularly useful when the data have been collected from various groups: unlike classical regression, which tries to fit a unique model for the whole dataset, a mixed effect model lets the regression coefficients vary group by group.

We will use mixed effect models to neutralise the effect of non-independent data, which are typical for diachronic linguistic data: the random effect is thus usually set as a diachronic variable, such as 'author'.

In R a mixed effect model for the response 'construction', predictors 'verb' and 'preposition', and random effect for 'author' is formulated as:

```
lmer(construction ~ verb + preposition + (1 |author), data = D, family =
```

    '(binomial')')
    In the case of a binary response, if we specify the binomial family, the default link function used is the logit (see section C.3.3).

## C.3.5 The model for the first investigation

The first investigation on Latin preverbs described in section 7.4.3 involved the use of mixed effect models to predict the value of the binary variable 'construction' based on the values of certain other variables.

In particular, we obtained the following model:

[^86]```
construction \(\sim\) prever \(b+e r a+\) case + animacy_relatum + figurative \(+(\) class \(\mid\) author \()\),
    family=binomial
```

As explained in section 7.4.3.1, in this model the response 'construction' (taking the two values 'CasePrev' and 'Preposition') is modeled on the variables 'preverb' ( $a b, a d$, circum, de, ex, in, intro-, ob, prae, sub, trans), 'era' (values 'Augustan', 'early_republic', 'imperial', 'late_imperial', 'late_republic' and 'medieval'), 'case' (accusative or ablative), 'animacy_relatum' ('animate' and 'non_animate') and 'figurative' ('figurative' and 'non_figurative'). 'Author' (values 'Caesar', 'Cicero', 'Jerome', 'Ovid', 'Petronius', 'Plautus', 'Propertius', 'Sallust', 'Thomas' and 'Vergil') and 'class' ('motion', 'rest' and 'transport') are the random effects.

Note that the response is binary and that the binomial distribution is chosen as the family distribution for the residuals. The syntax for the random effects refers to the fact that we allow the intercept to vary by the verb class and the slope by the author. In other words, instead of fitting a single model for the whole dataset by producing one intercept and one slope, we want to obtain different slopes and intercepts for the different authors and eras.

This section gives the details of the model we used, by presenting the graphical diagnostics for it (section C.3.5.1) and interpreting its output (section C.3.5.2).

## C.3.5.1 Graphical diagnostics

Before showing the coefficients given by the model, it is important to evaluate its goodness of fit to the data. Figure C. 4 is the residuals plot for the model (Gelman et al., 2009).

A binned plot shows the residuals of the model plotted against the estimated response. We will briefly explain these terms in our case; for a fuller illustration on how to interpret residuals plots, cf. Faraway (2006).

The $x$ axis of a binned plot displays the values for the response ('construction') as predicted by the model; the $y$ axis plots the corresponding average residuals from the model. The response, as it appears in the data, is binary, thus discrete, while the model provides the probabilities of the response having the value ' 0 ' $: 5$ this is why the $x$ axis displays continuous values between 0 and 1 .

If the model is a good model, then the residuals should be normally distributed and constant: in other words, there should not be any ascending or descending patterns in the binned plot. In addition, in a good regression model the variance should be constant; this translates into the fact that the grey line (which corresponds to the $95 \%$ confidence interval) should be as constant as possible. Finally, the points that fall inside the area delimited by the grey line are correctly assigned by the model, so it is desirable that most points fall into this area. Since good estimates are difficult with small datasets, in our case the number of points falling outside the area ( 3 out of 9 ) may be due to the small size of the dataset (containing 237 observations).

Given the fact that this was the model that gave the best fit to the data (based on the diagnostic plot in figure C.4), we considered it to be an acceptable model for our data.

[^87]

Figure C.4: Binned plot of the residuals of the linear model against fitted.

## C.3.5.2 Interpretation of the model

The details of the model are shown below, as they appear in the R output.

|  | Fixed effect | Est. coef. | Std. Error | $\operatorname{Pr}$ (preposition) | Effect size |
| ---: | :--- | ---: | ---: | ---: | ---: |
| 1 | (Intercept) | 1.45 | 0.63 | 0.81 | 0.81 |
| 2 | preverb.ad | -3.66 | 1.69 | 0.10 | 0.02 |
| 3 | preverb.de | -1.78 | 1.42 | 0.42 | 0.14 |
| 4 | preverb.ex | -1.13 | 0.93 | 0.58 | 0.24 |
| 5 | preverb.in | -2.59 | 1.42 | 0.24 | 0.07 |
| 6 | preverb.intro | 16.47 | 12722.08 | 1.00 | 1.00 |
| 7 | preverb.ob | -9.04 | 4.07 | $<0.01$ | $<0.01$ |
| 8 | preverb.sub | -2.79 | 1.50 | 0.21 | 0.06 |
| 9 | preverb.trans | -25.52 | 8050.37 | $<0.01$ | $<0.01$ |
| 10 | era.Augustan | -2.41 | 0.92 | 0.28 | 0.08 |
| 11 | era.imperial | 0.88 | 0.85 | 0.91 | 0.71 |
| 12 | era.medieval | 19.77 | 3244.85 | 1.00 | 1.00 |
| 13 | era.late_imperial | 19.64 | 3317.98 | 1.00 | 1.00 |
| 14 | era.late_republican | 0.82 | 0.75 | 0.91 | 0.69 |
| 15 | case.accusative | 2.46 | 1.46 | 0.98 | 0.92 |
| 16 | animacy_relatum.animate | -0.51 | 0.63 | 0.72 | 0.37 |
| 17 | figurative.figurative | 1.93 | 1.13 | 0.97 | 0.87 |

Table C.4: Estimated coefficient, standard error, probability of 'construction' and effect size of each predictor in the model.

Table C. 4 summarises the output details for each predictor value which deviates from the reference level, represented by the early republican era, the preverb $a b$-, the ablative case, inanimate relatum and non-figurative verb usage.

We will focus on the first two columns of table C.4, giving the estimated coefficient for each predictor's value and the standard error of this estimate. It is important that the error is smaller than the absolute value of the coefficient; otherwise the estimate is not reliable, as for 'preverb.intro', 'preverb.trans', 'era.medieval', 'era.late_imperial' and 'animacy_relatum.animate'. We will then focus on the rows whose names are in bold face in table C.4.

The estimated intercept is 1.45 (first row, first column in table C.4). Since this is a mixed effect model on a binary response and since we specified in the model that the residuals follow a binomial distribution, the default link function is the logit (cf. section C.3.3). Therefore, if we consider a coefficient (for example the estimated intercept 1.45) and calculate its value by the Inverse-logit function (0.81), this figure corresponds to the probability of 'construction' at the reference level.

Among the preverbs, $a d-$-, sub-, in-, $d e$ - and $e x$ - have a negative effect on 'construction' with respect to the reference level represented by $a b$-, which means that the probability of finding a preposition in the expression of relatum decreases if the preverb is one of those, compared with the probability of finding a preposition if the preverb is $a b-$.

The effect for the Augustan era is invlogit(-2.41)=0.08, which means that switching from the reference level given by the early republican era to the Augustan era causes a decrease (due to the minus sign in the coefficient) by $8 \%$ in the probability of 'construction' being 'preposition'; this probability is invlogit $(-2.41+1.45)=0.28$ when the era is

Augustan. All the other ages cause an increase in the probability of 'construction' $=$ 'preposition' with respect to the early republican age (reference): $71 \%$ (imperial) and $69 \%$ (late republican). The results for the medieval and the late imperial age have a high error because for these two eras there is a perfect overlap with the variable 'author'.

Finally, the accusative case governed by the preposition corresponding to the preverb causes an increase of $92 \%$ in the probability of finding a preposition, the animacy of the filler of the relatum a decrease of $37 \%$, and the figurative usage of the verb an increase of $13 \%$.

## C. 4 Correspondence Analysis

This section introduces Correspondence Analysis (CA; section C.4.1), a data analysis technique we used in the third study on Latin preverbs described in section 7.6. We outline the main ideas behind this technique (section C.4.2) and the procedure it follows (section C.4.3). Finally, we briefly introduce Multiple CA (MCA; section C.4.4) and give the details of our third study on Latin preverbs (section C.4.5).

## C.4.1 Introduction

CA is a technique in exploratory data analysis. It analyses tables of positive whole numbers (typically frequencies) to detect the correspondence between the rows (observations or individuals) and columns (variables or characteristics). Unlike traditional hypothesis testing techniques which aim at verifying hypotheses about relations between variables, exploratory data analysis aims at identifying systematic relations between variables and observations in the data without a priori expectations on these relations.

CA can also be defined as a multivariate descriptive technique for data analysis. In fact, it works on multidimensional data, i.e. sets of observations in the form of vectors in multidimensional spaces (see section B.5). Starting from such complex data, CA reduces the number of dimensions of the space: the output is a minimal space displaying the underlying structure of the data and providing a simplified version of the original input. As a matter of fact, CA is the only multidimensional method which allows us to study individuals and their characteristics in a unified representation. For example, clustering (see section B.5) only allows us to visualise the relation between individuals or between characteristics.

In addition, CA is a projection technique in that it projects the data points into a lower-dimensional space while preserving certain properties of the structure of the dataset as faithfully as possible. In this respect, CA is related to Factor Analysis.

Finally, CA is a graphical method, since it graphically displays row and column points onto a common space called biplot, which helps detect structural relationships between rows and columns in the data.

As we saw in section C.2, the $\chi^{2}$ test is able to detect an association between the rows and the columns of a matrix. However, it does not tell us where this association is concentrated, i.e. which are the significant individual associations between row-column pairs of the matrix. CA, on the other hand, provides a picture of these individual associations by a common representation in the biplot.

|  | verb |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | dico | do | eo | TOTAL |
| author | Caesar | 2 | 7 | 0 | 9 |
|  | Cicero | 3 | 1 | 0 | 4 |
|  | Planius | 2 | 0 | 11 | 13 |
|  | Sallust | 2 | 1 | 6 | 9 |
|  | TOTAL | 10 | 2 | 0 | 12 |
|  |  | 11 | 17 | 47 |  |

Table C.5: Frequency table of a constructed dataset with verbal tokens counted by verb lemma and author.

## C.4.2 The idea

The modern formulation of CA was made by Benzécri (1973) with reference to philological models regarding Chinese consonants and vowels. Later, other scholars followed Benzécri's idea and applied it to different areas: sociology, marketing, archeology, political sciences, linguistics, and other.

The main idea behind CA is to represent the data $K$ as the "sum" of a reduced model $K^{\prime}$, a residual unexplained variance $R$ and an error $E$ :

$$
K=K^{\prime}+R+E
$$

This resembles the approximation pursued by linear regression models (see section C.3). However, as Greenacre (2007, p. 47) points out, regression requires that one of the variables is considered as response; in contrast, CA does not require a response variable and treats all the variables at the same level.

## C.4.3 Procedure

In this section we will briefly outline the procedure common to all CA analyses. For a more in-depth discussion of each step, see Greenacre (2007).

Dataset In section B. 4 we showed the data format we usually deal with in linguistic analysis and how it can be turned into a frequency table. This table, or matrix, has in turn a geometrical interpretation if we consider its rows as vectors in a space defined by the number of columns.

For example, let us consider a study where we want to investigate the occurrences of Latin verbs in certain authors: we have a number of observations, corresponding to the verbal tokens, and two variables: 'author' and, say, 'verbal lemma'. Table C. 5 shows constructed frequencies for this case: rows correspond to authors and columns to verbal lemmas. CA can be applied to larger tables, but here we will make use of a toy example for the sake of clarity.

Profiles CA interprets the rows of the input table as points (called row profiles) in a multidimensional space defined by the columns, and it interprets the columns as points (called column profiles) in the space defined by the rows.

In the example above, this means considering the row points - 'Caesar', 'Cicero', 'Petronius', 'Plautus' and 'Sallust' - in the 3-dimensional space defined by the three

|  |  | verb |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | dico | do | eo | row masses |
| author | Caesar | 0.22 | 0.78 | 0 | 0.19 |
|  | Cicero | 0.75 | 0.25 | 0 | 0.09 |
|  | Plautus | 0.15 | 0 | 0.85 | 0.28 |
|  | Sallust | 0.22 | 0.11 | 0.67 | 0.19 |
|  | average row profile | 0.83 | 0.17 | 0 | 0.26 |
|  | 0.40 | 0.23 | 0.36 | 1 |  |

Table C.6: Matrix of row profiles from table C.5.
verbs dico, do, and eo; in addition, the three verbs are points in a 5 -dimensional space defined by the five author names.

Row and column profiles are calculated as the relative row and column frequencies, respectively. For example, the sum of the forth row in table C. 5 is $2+1+6=9$; hence, 'Plautus' corresponds to the row profile $(2,1,6) / 9=(0.22,0.11,0.67)$; similarly, the profile for dico is $(2,3,2,2,10) / 19=(0.10,0.16,0.10,0.10,0.05)$. We indicate the matrix of row profiles as $R$, represented in table C.6.

The average row profile is calculated as the row sums divided by the total sum. In other words, the average profile corresponds to the most expected behaviour. In the example, this is given by dividing the last row of table C. 5 by the grand total 47 : $(19,11,17) / 47=(0.40,0.23,0.36)$, this is the profile of all authors put together.

Masses The mass of a row $i$ is defined as the $i^{\text {th }}$ row total divided by the grand total; likewise for a column. For example, the mass of the first row in table C. 5 is $\frac{9}{47}=0.19$. Row masses are displayed in the last column of table C.6. The concept of mass helps us imagine the row and column profiles as weighted points in space: the profile space.

In the example above, the total frequency of the fourth row ('Plautus') is distributed among the verbs with the following weights: 0.22 (dico), 0.11 (do), 0.67 (eo) and 0.44 (facio). In other words, the weighted average or barycentre for 'Plautus' can be calculated by assigning weights of $\frac{2}{9}=0.22$ to dico, 0.11 to do and 0.67 to eo. If we associate the three verbs to the vertices of a triangle as in figure C.5, we can imagine that two cases (dico) are situated in the first vertex, one in the second (do) and six in the third (eo). Since eo has more cases, the average position of 'Plautus' is closer to the third vertex.

In the same way, each row profile (author) is positioned within the triangle as an average point and the profile values (i.e. the relative frequencies) correspond to weights associated to the vertices. These weights are the masses of the profiles. In addition, the average row profile is the barycentre of the five row profiles, as shown in figure C.5: it is centrally placed among the row profiles, but more attracted to the higher-frequency ones. For this reason, if the row profile points are close to the average row profile, it means that there is little variation in the table.

All that has been said regarding row profiles holds for column profiles too and the average column profile is the weighted average of the column profiles. In general, the average column profile coincides with the set of row masses, and the column masses are the elements of the average row profile.


Figure C.5: Geometrical visualisation of the row profiles from table C. 6 including the average profile.

|  |  | verb |  |  |
| :---: | :---: | :---: | :---: | :---: |
|  |  | dico | do | $e o$ |
| author | Caesar | 3.6 | 2.07 | 3.24 |
|  | Cicero | 1.6 | 0.92 | 1.44 |
|  | Plautus | 5.2 | 2.99 | 4.68 |
|  | Sallust | 3.6 | 2.07 | 3.24 |
|  |  | 2.76 | 4.32 |  |

Table C.7: Matrix of expected frequencies from table C. 5 under the hypothesis of independence.

Inertia and $\chi^{2}$-distance The hypothesis of independence in statistics assumes that there is no difference between the rows (or the columns) of a frequency table.

Let us consider the example of table C. 5 and assume that there is no preference in the way the authors used the verbs dico, do and $e o$ : in other words, the variables 'author' and 'verb' are independent. Since there are 9 occurrences of these verbs in Plautus and dico takes $40 \%$ of the total occurrences (see table C.6), we would expect $40 \%$ of the 9 occurrences in Plautus (i.e. $0.4 \cdot 9=3.6$ ) to be for dico. Similarly, we would expect $0.23 \cdot 9=2.07$ occurrences for $d o$ and $0.36 \cdot 9=3.24$ for eo in Plautus. If we calculate the relative expected and observed frequencies in this case we get: $(3.6,2.07,3.24) / 9=$ $(0.4,0.23,0.36)$ (expected) and $(2,1,6) / 9=(0.22,0.11,0.67)$ (observed). Note that the relative observed frequencies coincide with the row profiles.

Table C. 7 displays the expected frequencies from table C. 5 under the independence hypothesis. We may then ask the question: is the difference between expected and observed relative frequencies significant or not? In other words, is it likely that this difference arises by chance? Calculating the $\chi^{2}$ statistics answers this question. From
section C. 2 we recall:

$$
\begin{equation*}
\chi^{2}=\sum \frac{(\text { observed }- \text { expected })^{2}}{\text { expected }} \tag{C.5}
\end{equation*}
$$

If we use the formula (C.5) for the relative expected frequencies (table C.7) and the relative observed frequencies (table C.6) and we multiply each addend by the row total, we get

$$
\begin{align*}
\chi^{2}=\quad & \text { row total } \cdot \frac{(\text { observed row profile-expected row profile })^{2}}{\text { expected row profile }}=9 \cdot \frac{(0.22-0.40)^{2}}{0.40}+9 \cdot \frac{(0.78-0.23)^{2}}{0.23}+ \\
& +9 \cdot \frac{(0-0.36)^{2}}{0.36}+4 \cdot \frac{(0.75-0.40)^{2}}{0.40}+4 \cdot \frac{(0.25-0.23)^{2}}{0.23}+4 \cdot \frac{(0-0.36)^{2}}{0.36}+13 \cdot \frac{(0.15-0.40)^{2}}{0.40}+ \\
+ & 13 \cdot \frac{(0-0.23)^{2}}{0.23}+13 \cdot \frac{(0.85-0.36)^{2}}{0.36}+9 \cdot \frac{(0.22-0.40)^{2}}{0.40}+9 \cdot \frac{(0.11-0.23)^{2}}{0.23}+9 \cdot \frac{(0.67-0.36)^{2}}{0.36}+ \\
+ & 12 \cdot \frac{(0.83-0.40)^{2}}{0.40}+12 \cdot \frac{(0.17-0.23)^{2}}{0.23}+12 \cdot \frac{(0-0.36)^{2}}{0.36}=45.92 \tag{C.6}
\end{align*}
$$

The (total) inertia is defined as

$$
\begin{equation*}
\text { inertia }=\frac{\chi^{2}}{N} \tag{C.7}
\end{equation*}
$$

where $N$ is the grand total of the table. In the example, inertia $=\frac{45.92}{47}=0.97$. The total inertia measures how the individual profiles are spread around the weighted average profile, i. e. how much variance is in the table.

We can define a distance that measures how far each profile is from the average profile, thus accounting for how expected the behaviour of the profiles is. The $\chi^{2}$ distance between two row profiles $i_{1}$ and $i_{2}$ is defined as:

$$
\begin{equation*}
\chi^{2}\left(i_{1}, i_{2}\right)=\sum_{j=1}^{c} \frac{\left(R\left(i_{1}, j\right)-R\left(i_{2}, j\right)\right)^{2}}{A(j)} \tag{C.8}
\end{equation*}
$$

where $A(j)$ is the $\mathrm{j}^{\text {th }}$ coordinate of the average row profile and $R(i, j)$ is the $(i, j)^{\mathrm{th}}$ row profile. The $\chi^{2}$-distance can be seen as a weighted Euclidean distance (see section B.5.1, page 203 for a definition): in fact, the squared term corresponds to the coordinate of a point and it is weighted by the corresponding coordinate of the average profile. In the example, the $\chi^{2}$-distance between 'Plautus' and the barycentre is

$$
\begin{equation*}
\sqrt{\frac{(0.22-0.40)^{2}}{0.40}+\frac{(0.11-0.23)^{2}}{0.23}+\frac{(0.67-0.36)^{2}}{0.36}}=0.64 . \tag{C.9}
\end{equation*}
$$

The corresponding Euclidean distance would be

$$
\begin{equation*}
\sqrt{(0.22-0.40)^{2}+(0.11-0.23)^{2}+(0.67-0.36)^{2}}=0.38 . \tag{C.10}
\end{equation*}
$$

Note that, for example, the first term $(0.22-0.40)^{2}$ corresponds to the first coordinate of the row profile for 'Plautus' and is weighted by the inverse of the first coordinate of the average profile 0.40 in the formula for the $\chi^{2}$-distance.

The inertia in formula (C.7) is then given by the sum of the mass of each profile multiplied by the $\chi^{2}$-distance between the profile and the barycentre:

$$
\text { inertia }=\sum_{i=1}^{r}(i \text {-th mass }) \cdot\left(\chi^{2} \text { distance from } i \text {-th profile to barycentre }\right)^{2}
$$

We can obtain an analogous formula by referring to the column profiles.

A high inertia corresponds to large deviations of row profiles from the barycentre; the maximal inertia is reached when all the profiles lie at the vertices of the profile space. A special transformation of the coordinates allow us to interpret $\chi^{2}$-distances as Euclidean distances. For this, we need to replace the row profiles $R\left(i_{1}, j\right)$ and $R\left(i_{2}, j\right)$ in (C.8) with $\frac{R\left(i_{1}, j\right)}{\sqrt{A(j)}}$ and $\frac{R\left(i_{2}, j\right)}{\sqrt{A(j)}}$ respectively; $A(j)$ is the $j$-th coordinate of the average profile, thus always less than one. This transformation generally increases the coordinates of the points we are plotting, since we divide them by a number lower than one: this decrease is larger when $A(j)$ is small, i. e. when the $j$-th category is less frequent.

Dimensionality reduction The example in table C. 5 corresponds to a 3-dimensional space, which can be visualised. More often matrices for CA come with more columns (i.e. dimensions) so that it is necessary to reduce their dimensionality in order to get a manageable plot. This reduction happens at the expense of a loss of information: CA aims at limiting this loss as much as possible. For this purpose, CA searches for the optimal subspace $S$ that minimises the $\chi^{2}$-distance from the profile points. The idea behind this can be visualised in a simple example.

Figure C. 6 shows a two-dimensional space, i.e. a plane, with a line and four points: $a, b, c$ and $d$. Figure C. 7 shows a one-dimensional space, i.e. a line, with the projections of $a, b, c$ and $d$. The line covers most of the total variation of the points, thus it is a good representation of the original plane. However, not all the points are equally well projected: for instance, $b$ is an optimally represented point, since its projection in figure C. 7 coincides with its original position in figure C.6; $d$, instead, has the least precise projection.


Figure C.6: Four points in a two-dimensional space.
$S$ is obtained by using the Singular Value Decomposition (SVD), which finds the most "important" components of a matrix. More specifically, a matrix $A$ with $r$ rows and $c$ columns is decomposed by SVD into the product:

$$
A=U D V^{T}
$$

where $D$ is a diagonal matrix with the decreasing positive numbers $\lambda_{1}, \lambda_{2}, \ldots, \lambda_{k}$ in its diagonal and 0 everywhere else; $k$ is the rank of $A ;{ }^{6} V^{T}$ indicates the transpose of $V^{7}$

[^88]

Figure C.7: One-dimensional projection from figure C.6.
and $U$ and $V$ are so-called orthonormal matrices, i.e. $U U^{T}=I$ and $V V^{T}=I$, where $I$ is the identity matrix; ${ }^{8} \lambda_{1}, \lambda_{2}, \ldots, \lambda_{k}$ are called singular values.

If we define matrix $P$ as the original matrix $M$ divided by the grand total $N$, then SVD is applied to the matrix $R-I B$, since we center the space by the barycentre $B$. SVD decomposes this matrix as the product $U D V^{T}$ : where $D$ is a diagonal matrix with positive coefficients $\lambda_{1} \geq \lambda_{2} \geq \ldots \geq \lambda_{k}$ (singular values) and $U$ and $V$ are orthonormal matrices. The columns of $V$ are the orthonormal basis vectors that define $S$ and are called principal axes of inertia; the rows of matrix $F=N \cdot D$ are the projections of the row profiles onto $S$. SVD, hence, gives all the relevant information to project the original space into the optimal subspace.

Decomposition of inertia The total inertia of the matrix is decomposed into the direction of the principal axes of inertia. The first axis represents the direction where the inertia of the cloud is the maximum; the second axis maximises the inertia among all the directions orthogonal to the first axis, and so on. Geometrically, rows that contribute highly to a principal axis largely determine it.

It can be proven that the reduction of dimensionality of the row profiles and the reduction of dimensionality of the column profiles are solved by the SVD of the same matrix. In addition, the row analysis and the column analysis give the same decomposition of inertia into the principal axes. This explains the French name "analyse des correspondances" given to CA.

In the previous example, a CA on the data from table C. 5 gives a decomposition of inertia into $0.67(70.4 \%)$ and $0.28(29.6 \%)$ respectively for the two principal axes. This shows that, together, the two axes account for the whole inertia contained in the data. ${ }^{9}$

Biplot For $k=2$ or 3 , finding the optimal subspace $S$ allows us to plot the new coordinates in a two or three-dimensional space and get the CA of the row profiles. Then, the points are projected onto this subspace and constitute the final plot of the analysis.

Since both the row and the column analysis decompose the same inertia into the same system of principal axes, we can merge the two representations in one single geometric display which shows at the same time the projections of row and column points in the

[^89]subspace. Nevertheless, althought it is in general possible to interpret the distances between points in the same cloud (either row or column points), the interpretation of distances between points belonging to different clouds is more actively discussed.


Figure C.8: CA plot for the table C.5.
Figure C. 8 displays the plot of the CA for table C.5. Circles correspond to row points and triangles to column points. This plot is another version of figure C.5, since in this case the subspace coincides with the original space. Figure C. 8 shows three clear clusters of points and the similarity between the profiles for Plautus and Petronius on the one hand, and Cicero and Sallust on the other.

Diagnostics CA is also able to diagnose how faithful this new representation is with respect to the original profile space. In addition, it identifies what the points that contribute the most to this representation are. CA provides a system of diagnostic measures for each of the two dual spaces:

- contributions of the rows (and columns) to the axes, i. e. the inertia of the points projected onto the axes, which contributes to the principal inertia;
- contributions of the axes to the row (and column) points;
- quality of representation (cumulative sum of contributions of the axes for each point); this highlights well represented points.
We will give practical examples of how to interpret more complex CA plots and CA diagnostics in section C.4.5, while going into the actual results of our third study on Latin preverbs.


## C.4.4 Multiple Correspondence Analysis

MCA is the generalized extension of CA to handle more than two variables. The input to an MCA is a matrix whose rows correspond to observations and whose columns correspond to the values of the variables: the cell values of this matrix are ones or zeros, depending on whether the observation appears with the variable value or not.

The interpretation of the plot from an MCA is similar to that for a simple CA, as we will see in section C.4.5. In our analysis we performed an MCA by using the $m j c a$ function of the $c a$ package in R ( R Development Core Team, 2008) and setting the parameter lambda="adjusted". As pointed out by the authors in the package's manual (Greenacre and Nenadic, 2009), this means that the MCA is performed so that "the optimal scaling properties of MCA are conserved while raising the percentages of inertia". We plotted the results of the MCA analyses by setting the map parameter to 'symmetric', which plots the row and column points in principal coordinates.

## C.4.5 The second investigation

The second investigation on Latin preverbs - described in section 7.6 - consisted of an MCA of the second dataset collecting all four constructions for the realisation of the preverb's relatum; this analysis aimed at showing the interaction between the variable 'construction' and other relevant variables.

## C.4.5.1 Construction and author

At first, we considered the diachronic variable 'author' (see section 7.5.2.1). The singular values of an MCA on this dataset are reported in table C.8, from which we can see that the first two axes account for $88.6 \%$ of the total variance in the data.

| axis number | percentage | cumulated percentage |
| :---: | :---: | :---: |
| 1 | 66.5 | 66.5 |
| 2 | 22.1 | 88.6 |
| 3 | 11.4 | 100 |

Table C.8: Singular values of the MCA for the second investigation on the interaction between 'construction' and 'author'.

Figure C. 9 shows the two-dimensional plot of the analysis: the points are identified with triangles, whose size is proportional to their mass and colour intensity is proportional to their quality of representation. The first axis accounts for most of the inertia ( 0.24 , equal to $66.5 \%$ of the total inertia). Therefore, we will now concentrate on the horizontal layout of the plot.

If we project the points corresponding to the four constructions (as displayed in figure C.10, where the diagonal vector corresponds to the first axis), we get, from bottom left to top right, constructions 4 (PrepNonPrev) and 3 (PrepPrev) and then construction 2 (CaseNonPrev) and 1 (CasePrev). Constructions 3 (PrepPrev) and 4 (PrepNonPrev) are strongly associated with the Late Era authors (Jerome and Thomas), construction 1 (CasePrev) is strongly associated with Vergil and construction 2 is somehow shared by the other two poets (Ovid and Propertius) and Petronius and Sallust, although the latter ones behave close to expected, as Cicero and Plautus.


Figure C.9: Plot of the MCA for the second investigation on the interaction between 'construction' and 'era'.

If we now focus on the vertical axis - also displayed in figure C. 11 in the form of a diagonal vector - we see that this axis isolates the Early author (Plautus, left bottom of the figure C.11), strongly associated with construction 3 (PrepPrev), from the others.


Figure C.10: First axis of the MCA for the second investigation on the interaction between 'construction' and 'author'.

Table C.9, shows the diagnostics of the analysis: for each profile point (first column), it displays its mass (second column), the quality of its display in the subspace (third column), its inertia in permills of the total inertia (fourth column), its coordinates in each dimension (fifth and eight columns), the (relative) contribution of the principal axes to the point's inertia (sixth and ninth columns) and the (absolute) contribution of the point to the inertia of the axes in permills of the principal inertia (seventh and tenth columns).

The high mass and low inertia of the third construction and the Late era are confirmed by the values displayed in this table; similarly, we can see the high contribution of these points to the second axis and the high contribution of the other points to the first axis.

## C.4.5.2 More variables

We performed an MCA on the same dataset used in section C.4.5.1, but encoded with a higher number of variables: 'construction', 'era', 'preverb', 'animacy', and 'class', as illustrated in section 7.5.2.2. Table C. 10 contains the singular values, the explained inertia, the percentage of inertia inertia, and the cumulated percentage of inertia of this analysis; figure C. 12 shows the plot and table C. 11 displays the diagnostics.


Figure C.11: Second axis of the MCA for the second investigation on the interaction between 'construction' and 'author'.

| name | mass | qlt | inr | $\mathrm{k}=1$ | cor | ctr | $\mathrm{k}=2$ | cor | ctr |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| 1.CasePrev | 67 | 1811 | 184 | 936 | 1809 | 119 | 23 | 2 | 0 |
| 2.CaseNonPrev | 164 | 1458 | 68 | 254 | 891 | 22 | 154 | 567 | 14 |
| 3.PrepPrev | 169 | 3114 | 91 | -237 | 597 | 19 | -370 | 2516 | 82 |
| 4.PrepNoPrev | 101 | 2283 | 157 | -636 | 1486 | 83 | 354 | 797 | 45 |
| Caesar | 10 | 2990 | 8 | -116 | 97 | 0 | -482 | 2893 | 8 |
| Cicero | 24 | 758 | 26 | 188 | 190 | 2 | -247 | 569 | 5 |
| Jerome | 48 | 2118 | 118 | -795 | 1481 | 62 | 396 | 638 | 27 |
| Ovid | 21 | 2326 | 56 | 864 | 1608 | 33 | 439 | 718 | 15 |
| Petronius | 57 | 1711 | 13 | 259 | 1707 | 8 | -10 | 4 | 0 |
| Plautus | 95 | 3222 | 59 | 81 | 60 | 1 | -446 | 3161 | 67 |
| Propertius | 21 | 2249 | 44 | 797 | 1771 | 28 | 315 | 478 | 7 |
| Sallust | 41 | 1446 | 12 | 265 | 1383 | 6 | 43 | 63 | 0 |
| Thomas | 160 | 1709 | 76 | -370 | 1644 | 45 | 56 | 65 | 2 |
| Vergil | 23 | 2171 | 89 | 1131 | 1873 | 59 | 343 | 298 | 9 |

Table C.9: MCA Diagnostics of the column points for the second investigation on the interaction between 'construction' and 'author'. For each profile, the table contains its mass (second column), the quality of its display in the subspace (third column), its inertia in permills of the total inertia (fourth column), its coordinates in each dimension (fifth and eight columns), the (relative) contribution of the principal axes to the point's inertia (sixth and ninth columns) and the (absolute) contribution of the point to the inertia of the axes in permills of the principal inertia (seventh and tenth columns).

| axis number | inertia | percentage | cumulated percentage |
| :---: | :---: | :---: | :---: |
| 1 | 0.058 | 27.6 | 27.6 |
| 2 | 0.043 | 20.5 | 48.0 |
| 3 | 0.032 | 15.3 | 63.4 |
| 4 | 0.012 | 5.8 | 69.1 |
| 5 | 0.005 | 2.5 | 71.7 |
| 6 | 0.001 | 0.6 | 72.3 |
| 7 | $<0.0004$ | 0.1 | 72.4 |
| 8 | $<0.0001$ | 0.0 | 72.5 |
| 9 | $<0.0001$ | 0.0 | 72.5 |
|  | Total: 0.210 |  |  |

Table C.10: Singular values and inertia explained from the MCA for the second investigation on the interaction between 'construction', 'era', 'preverb', 'animacy', and 'class'.

## C.4.6 The third investigation

For the third investigation we worked on the following variables: 'preposition', 'era', type of complement ('compl'), and animacy of the lexical filler of the verb's complement ('a'), as illustrated on page 161.

The singular values of the MCA on this dataset are contained in table C.12, from which we can see that the first two axes account for $82.6 \%$ of the total variance in the data. Figure C. 13 shows the two-dimensional plot and table C. 13 displays the diagnostics of the analysis.


Figure C.12: Plot of the MCA for the second investigation on the interaction between 'construction', 'era', 'preverb', 'animacy', and 'class'.


Figure C.13: Plot of the MCA for the third investigation.

| name | mass | qlt | inr | $\mathrm{k}=1$ | cor | ctr | $\mathrm{k}=2$ | cor | ctr |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| 1CasePrev | 27 | 211 | 68 | -27 | 6 | 0 | 150 | 205 | 3 |
| 2CaseNonPrev | 66 | 2845 | 83 | -390 | 2392 | 41 | 157 | 453 | 8 |
| 3PrepPrev | 67 | 2849 | 47 | 161 | 748 | 7 | -251 | 2101 | 21 |
| 4PrepNoPrev | 40 | 1853 | 65 | 381 | 1793 | 24 | 65 | 60 | 1 |
| era.Classical | 79 | 1873 | 40 | -147 | 838 | 7 | 152 | 1036 | 9 |
| era.Early | 38 | 3281 | 59 | -95 | 115 | 1 | -463 | 3166 | 39 |
| era.Late | 83 | 1361 | 46 | 182 | 1176 | 11 | 67 | 185 | 2 |
| preverb.ab | 20 | 3270 | 70 | 79 | 36 | 1 | -696 | 3234 | 48 |
| preverb.ad | 57 | 1793 | 31 | -158 | 918 | 6 | -143 | 875 | 6 |
| preverb.circum | 2 | 648 | 20 | -523 | 623 | 3 | -97 | 25 | 0 |
| preverb.de | 12 | 3408 | 21 | 540 | 3317 | 14 | 83 | 91 | 0 |
| preverb.ex | 33 | 3081 | 34 | 357 | 2503 | 18 | 159 | 578 | 4 |
| preverb.in | 48 | 898 | 36 | -27 | 19 | 0 | 169 | 879 | 7 |
| preverb.intro | 3 | 371 | 22 | 319 | 262 | 1 | -190 | 108 | 0 |
| preverb.ob | 6 | 2958 | 21 | -703 | 2852 | 13 | 126 | 106 | 0 |
| preverb.prae | 4 | 1501 | 50 | -894 | 1213 | 13 | 405 | 288 | 3 |
| preverb.sub | 11 | 1957 | 24 | -160 | 227 | 1 | 410 | 1730 | 9 |
| preverb.trans | 4 | 545 | 40 | -0 | 0 | 0 | 490 | 545 | 5 |
| preverb.animate | 41 | 2559 | 48 | -345 | 2044 | 20 | -161 | 515 | 5 |
| preverb.non.animate | 159 | 2512 | 13 | 89 | 2001 | 5 | 42 | 511 | 1 |
| preverb.motion | 138 | 2800 | 30 | 147 | 1935 | 12 | 91 | 865 | 6 |
| preverb.rest | 25 | 2434 | 75 | -595 | 2354 | 37 | 102 | 79 | 1 |
| preverb.transport | 37 | 2731 | 57 | -138 | 241 | 3 | -410 | 2489 | 30 |

Table C.11: MCA Diagnostics of the column points for the second investigation on the interaction between 'construction', 'era', 'preverb', 'animacy', and 'class'. For each profile, the table contains its mass (second column), the quality of its display in the subspace (third column), its inertia in permills of the total inertia (fourth column), its coordinates in each dimension (fifth and eight columns), the (relative) contribution of the principal axes to the point's inertia (sixth and ninth columns) and the (absolute) contribution of the point to the inertia of the axes in permills of the principal inertia (seventh and tenth columns).

| nr | value | percentage | cumulated percentage |
| :---: | :---: | :---: | :---: |
| 1 | 0.033 | 79.5 | 79.5 |
| 2 | 0.001 | 3.2 | 82.6 |
|  | Total: 0.042 |  |  |

Table C.12: Singular values of the MCA for the third investigation.

## C. 5 Configural Frequency Analysis

In order to test Hypothesis 7.6.2.0.1, we are interested in analysing the relationship between the variables 'animacy' and 'preposition', which encodes the alternation between presence and absence of a preposition in the five constructions analysed.

In this section we describe some investigations we conducted in order to test this hypothesis: section C.5.1 illustrates the first attempts with MCA and regression models,

| name | mass | qlt | inr | $\mathrm{k}=1$ | cor | ctr | $\mathrm{k}=2$ | cor | ctr |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| no.preposition | 89 | 597 | 256 | 310 | 597 | 47 | -6 | 0 | 0 |
| preposition | 161 | 597 | 140 | -170 | 597 | 26 | 3 | 0 | 0 |
| eraClassical | 89 | 657 | 121 | 221 | 652 | 24 | $6-20$ | 5 | 1 |
| eraEarly | 55 | 379 | 31 | 32 | 32 | 0 | 105 | 347 | 17 |
| eraLate | 106 | 605 | 134 | -202 | 585 | 24 | -38 | 20 | 4 |
| complGoal | 174 | 589 | 36 | 78 | 523 | 6 | -28 | 66 | 4 |
| complSource | 76 | 589 | 82 | -177 | 523 | 13 | 63 | 66 | 8 |
| a.animate | 48 | 709 | 162 | 363 | 700 | 35 | 41 | 9 | 2 |
| a.non.animate | 202 | 709 | 38 | -86 | 700 | 8 | -10 | 9 | 1 |

Table C.13: MCA Diagnostics of the column points for the third investigation. For each profile, the table contains its mass (second column), the quality of its display in the subspace (third column), its inertia in permills of the total inertia (fourth column), its coordinates in each dimension (fifth and eight columns), the (relative) contribution of the principal axes to the point's inertia (sixth and ninth columns) and the (absolute) contribution of the point to the inertia of the axes in permills of the principal inertia (seventh and tenth columns).
while section C.5.2 gives an introduction to Configural Frequency Analysis (CFA) and finally section C.5.3 gives the details for a CFA on our data.

## C.5. 1 First attempts

The MCA exploration reported in sections 7.6 .2 and C.4.6 is able to capture the main patterns in the data, showing an association between inanimate fillers, the Late Era authors and the fourth and fifth construction (PrepNoPrev and PrepOther) on the one hand, and the earlier authors, the first and second construction (CasePrev and CaseNonPrev) and animate fillers on the other. We can imagine that the 'non-animate' point in figure 7.8 was attracted towards the point labeled as 'era.Late' because in the Late Era authors we find an unexpectedly high number of inanimate fillers. However, MCA is not able to detect associations appearing within specific subsets simply because its philosophy aims at detecting the overall structure in the data. If we want to investigate whether the earlier authors - while generally preferring non-prepositional constructions - used the prepositional constructions specifically with animate fillers, we need a different technique.

A first attempt would involve a regression model which predicts the value of the binary variable 'preposition' based on the animacy of the filler and a diachronic variable ('era' or 'author'), with a possible random effect for 'author' or 'era', respectively. In this case we would obtain a coefficient for each value of the predictors ('animate', 'no.animate', and the authors or eras); these coefficients would tell us the effect of each of these variable values on the outcome given by the variable 'preposition'. Again, this would be a global effect, which does not provide insights on the authors' preference for animate fillers when they use a preposition. On the other hand, fitting a model for each single diachronic category makes the analysis poorer, because it removes variability from the dataset, which is exactly the type of input that is best dealt with by regression models.

For these reasons we tried fitting a regression model with interaction effects between 'animacy' and 'author' (or 'era').

$$
\text { preposition } \sim \text { animacy } * \text { era }+(1 \mid \text { author }), \text { family=binomial }
$$

This way, the output would give us a coefficient for each value of the variables and also for each combination of them. Such models do not fit the data well enough, which means that the presence vs. absence of preposition in the data cannot be explained by using these predictors and these random effects.

## C.5.2 Configural Frequency Analysis

Configural Frequency Analysis (CFA; von Eye 1990) is an exploratory method that deals with tables recording frequency of pattern co-occurrences. CFA aims at finding those patterns in the data (called configurations) that are significantly more frequent or significantly less frequent than expected by chance: the former ones are called types and the latter ones antitypes.

The $\chi^{2}$ version of CFA can be thought of as a multidimensional generalization of the $\chi^{2}$ test to cases where more than two dimensions are involved. In fact, it assigns a $p$-value to each configuration, thus identifying which configurations behave in a significantly different way from expected.

In section C.5.3 we will summarise the way CFA works by describing the actual data we conducted the analysis on.

## C.5.3 A Configural Frequency Analysis for Direct Object Marking

Table C. 14 contains the frequencies of the different combinations of the values of the variables 'era', ${ }^{10}$ 'animacy' and 'preposition' in our data; the rows of this table are the configurations. ${ }^{11}$

| animacy | preposition | era | frequency | expected frequency |
| :---: | :---: | :---: | :---: | :---: |
| animate | no.preposition | Classical | 27 | 9.61 |
| non.animate | preposition | Late | 126 | 88.29 |
| non.animate | no.preposition | Late | 25 | 48.44 |
| animate | preposition | Late | 10 | 20.84 |
| animate | preposition | Classical | 8 | 17.51 |
| non.animate | preposition | Classical | 58 | 74.18 |
| animate | no.preposition | Early | 10 | 5.89 |
| non.animate | no.preposition | Classical | 49 | 40.70 |
| animate | no.preposition | Late | 8 | 11.43 |
| animate | preposition | Early | 13 | 10.73 |
| non.animate | no.preposition | Early | 22 | 24.94 |
| non.animate | preposition | Early | 42 | 45.45 |

Table C.14: Frequency counts and expected frequencies of the different combinations of the values of 'era', 'animacy' and 'preposition' in our data.

[^90]If we assume that the variables are independent we get a random distribution, which gives the expected frequency of each configuration. CFA compares these expected frequencies with the observed frequencies and assigns $p$-values according to different statistical tests. If the expected frequency of a configuration is lower than the observed frequency then the configuration is a type, otherwise it is either an antitype (higher) or not classified. For example, in table C. 14 the first configuration (first row) is a type.

A $\chi^{2}$ test with a Bonferroni correction gave a significant result at the 0.05 level for the first three rows in table C.14. This result says that the only patterns showing a significant deviance from the expected behaviour are absence of preposition + animate filler in the Classical authors (occurring more frequently than expected), presence of preposition + inanimate fillers and absence of preposition + inanimate fillers for the Late era authors. This confirms the results from the MCA investigations (cf. sections 7.6 and C.4.6). On the other hand, the configurations animate + preposition + Classical or Early - which would be the interesting ones in Hypothesis 7.6.2.0.1 - do not display a significant $p$-value.

In conclusion, only three patterns of co-occurrence for 'animacy', 'preposition' and 'era' behave in a significantly different way from expected. This also explains why a regression model with interaction effects failed to fit the data: it is not possible to predict the value of 'preposition' as a linear function of 'animacy' and 'era' and their interactions (up to a random error) because for most cases this association is not systematic enough to be captured.

## Bibliography

## Primary sources

Aquinas, T. (1856-1858). Sancti Thomae Aquinatis, doctoris angelici, Ordinis praedicatorum Commentum in quatuor libros Sententiarum magistri Petri Lombardi, adjectis brevibus adnotationibus. Parma: Fiaccadori.

Caesar, C. J. (1914). Commentarii Rerum in Gallia Gestarum VII. A Hirti Commentarius VIII. Oxford: Clarendon.

Cicero, M. T. (1909). M. Tulli Ciceronis Orationes: Recognovit brevique adnotatione critica instruxit Albertus Curtis Clark. Albert Clark. Oxonii. e Typographeo Clarendoniano. Scriptorum Classicorum Bibliotheca Oxoniensis.

Crispi, C. S. (1919). Catilina, Iugurtha, Orationes et epistulae excerptae de historiis. Recognovit Axel W. Ahlberg. Leipzig: Teubner.

Greenough, J. B. (1882). Bucolica, Aeneis, Georgica. The Greater Poems of Virgil. Boston: Ginn \& Co.

Jerome. Vulgate. Bible Foundation and On-Line Book Initiative.
Ovid (1892). Metamorphoses. Gotha: Friedr. Andr. Perthes.
Petronius (1913). Satyricon. London: William Heinemann.
Plautus, T. M. (1895). Plauti Comoediae. Berlin: Weidmann.
Propertius (1995). Charm. Los Angeles: Sun and Moon Press.

## Dictionaries and Thesauri

(2002). Thesaurus Linguae Latinae, CD-ROM. Thomas technology solutions.

Bianco, M. T. (1996). Valenzlexikon Deutsch-Italienisch / Dizionario della valenza verbale. Heidelberg: Groos.

Blumenthal, P. and G. Rovere (1998). Wörterbuch der italienischen Verben. Konstruktionen, Bedeutungen, Übersetzungen. Stuttgart: Klett.

Calonghi, F. (1967). Dizionario latino-italiano. Torino: Rosenberg \& Sellier.
Engel, U. and E. Savin (1983). Valenzlexikon deutschrumänisch. Dicionar de valen german-román. Heidelberg.

Forcellini, E. (1940). Lexicon totius latinitatis (fifth ed.). Patavii.
Glare, P. G. W. Oxford Latin Dictionary.
Herbst, T., D. Heath, I. F. Roe, and D. Götz (2004). A valency Dictionary of English. A corpus-based analysis of the Complementation Patterns of English Verbs, Nouns and Adjectives. Berlin/New York: Mouton de Gruyter.
Lewis, C. T. and C. Short (1879). A Latin Dictionary, Founded on Andrews' edition of Freund's Latin dictionary revised, enlarged, and in great part rewritten by Charlton T. Lewis, Ph.D. and Charles Short, LL.D. Oxford: Clarendon Press. url: http://www.lib.uchicago.edu/efts/PERSEUS/Reference/lewisandshort.html.

## References

Abney, S. and M. Light (1999). Hiding a semantic hierarchy in a Markov Model. In Proceedings of the ACL Workshop on Unsupervised Learning in Natural Language Processing. University of Maryland, College Park, pp. 1-8.

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

Agirre, E., A. Olatz, D. Martinez, and E. Hovy (2000). Enriching very large ontologies using the WWW. In Proceedings of the Ontology Learning Workshop, organized by ECAI, Berlin, Germany.

Alishahi, A. (2008). A probabilistic model of early argument structure acquisition. Ph. D. thesis, Department of Computer Science, University of Toronto.

Allen, A. (1981). The development of prefixal and parasynthetic verbs in Latin and Romance. Romance Philology XXXV(1), 79-88.

Alonso, L., I. Castellón, and N. Tinkova Tincheva (2007). Obtaining coarse-grained classes of subcategorization patterns for Spanish. In Proceedings of the International Comference Recent Advances in Natural Language Processing (RANLP 2007), Borovets, Bulgaria.

Apollon, D. (2001-2003). Analytica (C).
Attardi, G. (2006). Experiments with a multilanguage non-projective Dependency Parser. In Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X). New York City, pp. 166-170. Association for Computational Linguistics.

Baayen, R. H. (2008). Analyzing Linguistic Data. Cambridge: Cambridge University Press.

Baker, C., C. Fillmore, and J. Lowe (1998). The Berkeley FrameNet project. In Proceedings of COLING-ACL 1998. Montreal.

Baldi, P. (1979). Typology and the Indo-European prepositions. Indogermanische Forschungen 84, 49-61.

Bamman, D. and G. Crane (2006). The design and use of a Latin dependency treebank. In J. Hajic and J. Nivre (Eds.), Proceedings of the Fifth International Workshop on Treebanks and Linguistic Theories (TLT 2006). Prague, pp. 67-78.

Bamman, D. and G. Crane (2007). The Latin Dependency Treebank in a cultural heritage digital library. In Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007). Prague, pp. 33-40.

Bamman, D. and G. Crane (2008). Building a dynamic lexicon from a digital library. In Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2008). Pittsburgh.

Bamman, D., G. Crane, M. Passarotti, and S. Raynaud (2007a). A collaborative model of treebank development. In Proceedings of the Sixth Workshop on Treebanks and Linguistic Theories (TLT 2007). Bergen, pp. 1-6.

Bamman, D., G. Crane, M. Passarotti, and S. Raynaud (2007b). Guidelines for the syntactic annotation of Latin treebanks. Technical report, Tufts Digital Library.

Bamman, D., M. Passarotti, R. Busa, and G. Crane (2008). The annotation guidelines of the Latin Dependency Treebank and Index Thomisticus Treebank. The treatment of some specific syntactic constructions in Latin. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008). Marrakech.

Bamman, D., M. Passarotti, and G. Crane (2008). A case study in treebank collaboration and comparison: accusativus cum infinitivo and subordination in Latin. Prague Bulletin of Mathematical Linguistics 90, 109-122.

Basili, R., M. Pazienza, and M. Vindigni (1997). Corpus-driven unsupervised learning of verb subcategorization frames. In Proceedings of the 5th Congress of the Italian Association for Artificial Intelligence on Advances in Artificial Intelligence, London, UK, pp. 159-170. Springer-Verlag.

Bentivogli, L. and E. Pianta (2004). Extending WordNet with Syntagmatic Information. In Proceedings of the Second Global WordNet Conference, Brno, Czech Republic, pp. 47-53.

Benveniste, E. (1949). Le système sublogique des prépositions en latin. Travaux du Cercle Linguistique de Copenhague V. Recherches structurales, 177-184.

Benzécri, J.-P. (1973). L’Analyse des Données, Volume 1. Dunod.
Bilisoly, R. (2008). Practical text mining with Perl. Wiley.
Bloomfield, L. (1933). Language. New York: Holt.
Böhmová, A., J. Hajič, E. Hajičová, and B. Hladká (2003). The Prague Dependency Treebank: a three-level annotation scenario. In A. Abeillé (Ed.), Treebanks: building and using parsed corpora, pp. 103-128. Dordrecht: Kluwer Academic Publishers.

Booij, G. and A. van Kemenade (2003). Preverbs: an introduction. Yearbook of Morphology, 1-11.

Bortolussi, B. (2005). Le double accusatif aver le verbes préfixés en circum- et trans-. In C. Moussy (Ed.), La composition et la préverbation en latin, pp. 277-292. Presses de l'Université Paris-Sorbonne.

Bossong, G. (1983). Animacy and markedness in universal grammar. Glossologia 2-3, 7-20.

Brachet, J. P. (2000). Recherches sur les préverbes de- et ex- du latin. Bruxelles: Latomus. Revue d'études Latines.

Brachet, J.-P. (2005). Préverbes en intro en cours de constitution chez Plaute et Térence. In C. Moussy (Ed.), La composition et la préverbation en latin, pp. 309-320. Presses de l'Université Paris-Sorbonne.

Brent, M. R. (1991). Automatic acquisition of subcategorization frames from untagged text. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, pp. 209-214.

Brent, M. R. and R. C. Berwick (1991). Automatic acquisition of subcategorization frames from tagged text. In Proceedings of the workshop on Speech and Natural Language at the Human Language Technology Conference, Pacific Grove, California, pp. 342-345.

Brinton, L. J. (1988). The development of English Aspectual Systems. Aspectualizers and Post-Verbal Particles. Cambridge: Cambridge University Press.

Brinton, L. J. and E. C. Traugott (2005). Lexicalization and language change. Cambridge: Cambridge University Press.

Briscoe, T. and J. Carroll (1997). Automatic extraction of subcategorization from corpora. In Proceedings of the Fifth Applied Natural Language Processing Conference, pp. 356-363.

Buchholz, S. and E. Marsi (2006). CoNLL-X shared task on multilingual dependency parsing. In Proceedings of the X CoNLL Shared Task, pp. 149-164.

Bühler, K. (1934). Sprachtheorie: die Darstellungs-funktion der Sprache. Jena: Gustav Fischer, Stuttgart.

Busa, R. (1974-1980). Index Thomisticus: sancti Thomae Aquinatis operum omnium indices et concordantiae, in quibus verborum omnium et singulorum formae et lemmata cum suis frequentiis et contextibus variis modis referuntur quaeque / consociata plurium opera atque electronico IBM automato usus digessit Robertus Busa SJ. Stuttgart - Bad Cannstatt: Frommann - Holzboog.

Buysschaert, J. (1982). Criteria for the Classification of English Adverbials. Brussel: AWLSK.

Calboli, G. (2005). Composition avec le préfixe privatif in- chez Horace et les poètes de son temps. In C. Moussy (Ed.), La composition et la préverbation en latin, pp. 71-87. Presses de l'Université Paris-Sorbonne.

Calvo, H. and A. Gelbukh (2004). Acquiring selectional preferences from untagged text for prepositional phrase attachment disambiguation. Lecture notes in computer science 3136, 207-216.

Campbell, L. (2001). What's wrong about grammaticalization? Language Sciences 23, 93-112.

Carnie, A. (2007). Syntax. A Generative Introduction (2 ${ }^{\text {nd }}$ ed.). Blackwell.
Carroll, J., T. Briscoe, and A. Sanfilippo (1998). Parser evaluation: a survey and a new proposal. In Proceedings of the First International Conference on Language Resources and Evaluation (LREC 1998). Granada, pp. 447-454.

Carroll, J., G. Minnen, and T. Briscoe (1998). Can subcategorization probabilities help a statistical parser? In Proceedings of the 6th ACL/SIGCDAT Workshop on Very Large Corpora. Montreal, pp. 149-164.

Carvalho, P. d. (1983). Le système des cas latins. Aperçu d'une théorie morphosémantique. In H. Pinkster (Ed.), Latin Linguistics and Lingusitic Theory. Proceedings of the First International Colloquium on Latin Linguistics. Amsterdam, April 1981, Amsterdam, pp. 59-71. Benjamins.

Carvalho, P. d. (1996). Cas et préposition en linguistique latine et en linguistique théorique. In H. Rosén (Ed.), Aspects of Latin. Papers from the Seventh International Colloquium on Latin Linguistics, Jerusalem, April 1993, Innsbruck. Innsbrucker Beiträge zur Sprachwissenschaft.

Carvalho, P. d. (1998). La grammaire de la préverbation ou: préverbation et "ergativité". In B. García-Hernández (Ed.), Estudios de Lingüística latina. Actas del IX Coloquio Internacional de Lingüística Latina. Universidad Autonóma de Madrid, 1997, pp. 273-289. Ediciones Clásicas.

Castellón, I., A. Fernández-Montraveta, G. Vázquez, L. Alonso Alemany, and J. A. Capilla (2006). The SenSem Corpus: a Corpus Annotated at the Syntactic and Semantic Level. In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), pp. 355-358.

Castiglioni, L. and S. Mariotti (1993). IL vocabolario della lingua latina. Loescher.
Chomsky, N. (1965). Aspects of the Theory of Syntax. Cambridge, MA: MIT Press.
Chomsky, N. (1970). Remarks on nominalization. In R. P. Jacobs, Roderick (Ed.), Reading in English Transformational Grammar, pp. 184-221. Waltham: Ginn.

Chomsky, N. (1981). Lectures on Government and Binding. Dordrecht: Foris.
Ciaramita, M. and M. Johnson (2000). Explaining away ambiguity: Learning verb selectional preference with bayesian networks. In Proceedings of the 18th conference on Computational linguistics. Saarbrücken, pp. 187-193.

Clark, S. 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, College Park, pp. 258-265.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences ( $2^{\text {th }}$ ed.). Hillsdale, NJ: Lawrence Erlbaum.

Coleman, R. (1991). Latin prepositional syntax in Indo-European perspective. In R. Coleman (Ed.), New Studies in Latin Linguistics. Selected papers from the 4 th International Colloquium on Latin Linguistics. Cambridge, 1987, Cambridge, UK, pp. 323-338. John Benjamins.

Collins, M. (1997). Three generative, lexicalised models for statistical parsing. In Proceedings of the 35th Annual Meeting of the ACL. Madrid, pp. 16-23.

Collins, M. J. (1996). A new statistical parser based on bigram lexical dependencies. In Proceedings of the 34th annual meeting on Association for Computational Linguistics, Morristown, pp. 184-191. Association for Computational Linguistics.

Covington, M. A. (1990). A dependency parser for variable-word-order languages. Technical report, Artificial Intelligence Programs, University of Georgia.

Craig, C. and K. Hale (1988). Relational preverbs in some languages of the Americas: typological and historical perspectives. Language $64(2), 312-344$.

Crane, G. (1991). Generating and parsing Classical Greek. Literary and Linguistic Computing 6(4), 243-245.

Crocco Galèas, G. and C. Iacobini (1992). Parasintesi e doppio stadio derivativo nella formazione verbale del latino. Archivio Glottologico Italiano 77, 167-199.

Croft, W. (2000). Explaining language change: An evolutionary approach. London: Longman.

Cuzzolin, P. (1995). A proposito di sub vos placo e della grammaticalizzazione delle adposizioni. Archivio Glottologico Italiano 80, 122-143.
Cuzzolin, P., I. Putzu, and P. Ramat (2006). The Indo-European Adverb in diachronic and typological perspective. Indogermanische Forschungen 111, 1-38.

Dardano, M. (1978). La formazione delle parole nell'italiano di oggi. Roma: Bulzoni.
De Vivo, A. (1994). Lingua e comico in Plauto. Bollettino di studi latini 24 (2), 417-431.
Dell'Orletta, F., M. Federico, E. Giovannetti, A. Lenci, S. Marchi, A. Trabucco, and V. Pirrelli (2007). Specifiche di Named Entity Recognition per l'Italiano. Technical report, ILC_TW_2007_06-Progetto Smartext.

Dempster, A., L. N. M. and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society 39(B), 1-38.

Denison, D. (2003). Log(ist)ic and simplistic S-curves. In R. Hickey (Ed.), Motives for language change, pp. 54-70. Cambridge: Cambridge University Press.

Dikken, M. d. (1995). Particles: On the Syntax of Verb-Particle, Triadic, and Causative Constructions. Oxford: Oxford University Press.

Dowty, D. (1991). Thematic proto-roles and argument selection. Language 67, 547-619.
Dufresne, M., F. Dupuis, and M. Tremblay (2003). Preverbs and particles in Old French. Yearbook of Morphology, 33-59.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. Computational Linguistics 19, 61-74.

Durante, M. (1981). Dal latino all'italiano moderno. Saggio di storia linguistica e culturale. Bologna: Zanichelli.

Echarte Cossío, M. J. (1998). Preposiciones y casos en latín: propuesta de un sistema conjuncto. In B. García-Hernández (Ed.), Estudios de Lingüística latina. Actas del IX Coloquio Internacional de Lingüística latina. Universidad Autonóma de Madrid, 1997, pp. 309-323. Ediciones Clásicas.

Erk, K. (2007). A simple, similarity-based model for selectional preferences. In Proceedings of the 45 th Annual Meeting of the Association for Computational Linguistics. Prague, pp. 216-223.

Ernout, A. and F. Thomas (1972). Syntaxe latine. Paris: Klincksieck.
Evert, S. (2006). How Random is a Corpus? The Library Metaphor. Zeitschrift für Anglistik und Amerikanistik 54(2), 177-190.

Evert, S. and M. Baroni (2006). zipfR: Statistical models for word frequency distributions. R package version 0.6-0.

Faraway, J. J. (2006). Extending the linear model with R. Boca Raton, FL: Chapman \& Hall/CRC.

Fillmore, C. (1968). The test for case. In E. Bach and R. T. Harms (Eds.), Universals in Linguistic Theory, pp. 1-88. New York: Holt, Rinehart and Winston.

Fillmore, C. (1985). Syntactic Intrusions and the Notion of Grammatical Construction. Berkeley Linguistic Society 11, 73-86.

Fillmore, C. (2007). Valency issues in FrameNet. In T. Herbst and K. Götz-Votteler (Eds.), Valency. Theoretical, Descriptive and Cognitive Issues, Trends in Linguistics, pp. 129-161. Mouton de Gruyter.

Fillmore, C., C. R. Johnson, and M. R. L. Petruck (2003). Background to Framenet. International Journal of Lexicography 16(3), 235-250.

Fillmore, C. J. (1995). Constituency vs. dependency. In F. Madray-Lesigne and J. Richard-Zappella (Eds.), Lucien Tesnière aujourd'hui. Actes du Colloque International C. N. R. S. URA 1164, Univesrité de Rouen, 1991, Louvain, pp. 93-104. Société pour l'Information Grammaticale: Éditions Peeters.

Firth, J. R. (1957). Studies in Linguistic Analysis, Chapter A synopsis of linguistic theory 1930-1955. Philological Society.

Fischer, I. (1971). Encore sur le caractère de la langue de Plaute. Studclas 13, 59-78.
Fischer, O. (2007). Morphosyntactic change. Functional and formal perspectives. Oxford surveys in syntax and morphology. Oxford: Oxford University Press.

Francis, W., . K. H. (1979). Manual of Information to Accompany a Standard Corpus of Present-Day Edited American English, for Use with Digital Computers (Revised and amplified ed. ed.). Rhode Island: Department of Linguistics, Brown University.

Fraser, N. (1990). Prolegomena to a formal theory of dependency grammar. UCL Working Papers in Linguistics 2, 298-319.

Fruyt, M. (1998). La grammaticalisation en latin. In B. García-Hernández (Ed.), Estudios de Lingüística latina. Actas del IX Coloquio Internacional de Lingüística latina, Universidad Autonóma de Madrid, 1997, pp. 877-889. Ediciones Clásicas.

Gaide, F. (2005). À propos des préverbes en com-; "couper" et "broyer" dans le $D e$ Medicamentis de Marcellus. In C. Moussy (Ed.), La composition et la préverbation en latin, pp. 263-276. Presses de l'Université Paris-Sorbonne.

Gale, W. A. and G. Sampson (1995). GoodTuring frequency estimation without tears. Journal of Quantitative Linguistics 2, 217-237.

García-Hernández, B. (2004). Synonymie et analyse fonctionnelle dans le système préverbal latin. Revue des études latines 72, 25-38.

García-Hernández, B. (2005). L'antonymie aspectuelle des préverbes allatifs et ablatifs. In C. Moussy (Ed.), La composition et la préverbation en latin, pp. 229-241. Presses de l'Université Paris-Sorbonne.

Garcìa Jurado, F. (1991). Los sintagmas preposicionales $e x, a b, d e+$ ablativo en latìn clasico: sistema semàntico. Minerva 5, 189-206.

Geeraerts, D. (2006). Methodology in cognitive linguistics. In G. Kristiansen, M. Achard, R. Dirven, and F. J. Ruiz de Mendoza Ibáñez (Eds.), Cognitive linguistics: Current applications and future perspectives, pp. 21-50. Berlin: Mouton de Gruyter.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2004a). Bayesian Data Analysis (2 $2^{\text {th }}$ ed.). Boca Raton: Chapman \& Hall/CRC.

Gelman, A., Y.-S. Su, M. Yajima, J. Hill, M. G. Pittau, J. Kerman, and T. Zheng (2009). arm: Data Analysis Using Regression and Multilevel/Hierarchical Models. R package version 1.2-9.

Giacomelli, R. (1993). Storia della lingua latina. Roma: Jouvence.
Goggi, Sara, e. a. (1997). Italian Corpus Documentation. Technical Report WP2.11, ILC, Pisa.

Goldberg, A. (2006). Constructions at Work: The Nature of Generalization in Language. Oxford: Oxford University Press.

Goldberg, A. E. (1995). Constructions: a construction grammar approach to argument structure. The University of Chicago Press.

Gorard, S. (2003). Quantitative methods in social science. New York/London: Continuum.

Götz-Votteler, K. (2007). Describing semantic valency. In T. Herbst and K. GötzVotteler (Eds.), Valency. Theoretical, Descriptive and Cognitive Issues, Trends in Linguistics, pp. 37-49. Mouton de Gruyter.

Greenacre, M. (2007). Correspondence Analysis in practice (2 $2^{\text {th }}$ ed.). Boca Raton, FL: Chapman \& Hall/CRC.

Greenacre, M. and O. Nenadic (2009). ca: Simple, Multiple and Joint Correspondence Analysis. R package version 0.3-2.

Griffe, M. (1995). Théorie de la valence et construction du verbe Latin. In F. MadrayLesigne and J. Richard-Zappella (Eds.), Lucien Tesnière aujourd'hui. Actes du Colloque International C. N. R. S. URA 1164. Université de Rouen, 1992, Louvain, pp. 269-276. Société pour l'Information Grammaticale: Éditions Peeters.

Gropen, J., S. Pinker, M. Hollander, R. Goldberg, and R. Wilson (1989). The learnability and acquisition of the dative alternation in English. Language 65(2), 203-257.

Gruber, J. (1976). Studies in Lexical Relations. Ph. D. thesis, MIT.
Haegeman, L. (1996). Introduction to Government and Binding Theory (2 $2^{\text {nd }}$ ed.). Blackwell.

Hajič, J., J. Panevova, Z. Uresova, and A. Bemova (1999). Annotations at analytical level: Instructions for annotators (English translation by Z. Kirschner). Technical report, UFAL MFF UK, Prague.

Hajič, J., J. Panevová, Z. Urešová, A. Bémová, V. Kolárová-Reznícková, and P. Pajas (2003). PDT-VALLEX: Creating a large coverage valency lexicon for treebank annotation. In J. Nivre and E. Hinrichs (Eds.), Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT 2003). Växjö, Volume 9, pp. 57-68. Växjö University Press.

Halácsy, P., A. Kornai, and C. Oravecz (2007). HunPos an open source trigram tagger. In Proceedings of the ACL 2007 Demo and Poster Sessions, pp. 209-212.

Hamp, B. and H. Feldweg (1997). GermaNet - a Lexical-Semantic Net for German. In Proceedings of the ACL Workshop on Automatic Information Extraction and Building Lexical Semantic Resources for NLP Applications. Madrid.

Happ, H. (1976a). Grundfragen einer Dependenz-Grammatik des Lateinischen. Goettingen. Vandenhoeck \& Ruprecht.

Happ, H. (1976b). Syntaxe latine et théorie de la valence: essai d'adaptation au latin des théories de lucien tesnière. Langages Linguistique et latin(50), 51-72.

Harabagiu, S. M. and D. I. Moldovan (2000). Enriching the WordNet taxonomy with contextual knowledge acquired from text. pp. 301-333.

Harrell, F. E., K. L. Lee, and D. B. Mark (1996). Multivariate prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in Medicine 15(4), 361-387.

Haug, D. (2009). Grammatical Change in Indo-European Languages, Chapter Does Homeric Greek have prepositions? Or local adverbs? (And what's the difference anyway?), pp. 103-120. John Benjamins Publishing Company.

Haug, D. T. T. and M. L. Jøndal (2008). Creating a parallel treebank of the old IndoEuropean Bible translations. In Proceedings of Language Technologies for Cultural Heritage Workshop (LREC 2008). Marrakech, pp. 27-34.

Haverling, G. (2000). On SCO- verbs, prefixes and semantic functions. Number 64 in Studia Graeca et Latina Gothoburgensia. Göteborg: Acta Universitatis Gothoburgensis.

Heine, B., U. Claudi, and F. Hünnemeyer (1991). Grammaticalization. A conceptual framework. Chicago/London: The University of Chicago Press.

Heine, B. and T. Kuteva (2002). World lexicon of grammaticalization. Cambridge: Cambridge University Press.

Helbig, G. (1992). Probleme der Valenz- und Kasustheorie. Tübingen: Niemeyer.
Hellwig, P. (1986). Dependency unification grammar. In Proceedings of the 11th International Conference on Computational Linguistics. Bonn, pp. 195-198.

Herbst, T., D. Heath, and H.-M. Dederding (1980). Grimm's grandchildren: current topics in German linguistics. Longman.

Heslin, T. P. (1987). Prefixation in Latin. Lingua 72, 133-154.
Hindle, D. (1990). Noun classification from predicate argument structures. In Proceedings of ACL 1990, Pittsburg.

Hinrichs, E. W. and H. Telljohann (2009). Constructing a valence lexicon for a treebank of German. In Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories (TLT 7). Groningen, pp. 41-52.

Hinton, P. R. (2004). Statistics explained. Routledge.
Hlaváčková, D., H. A. and V. Kadlec (2006). Text, Speech and Dialogue, Chapter Exploitation of the VerbaLex Verb Valency Lexicon in the Syntactic Analysis of Czech. Berlin/Heidelberg: Springer.

Hofmann, J. B. H. (1980). La lingua d'uso latina. Bologna: Pàtron.
Hopper, P. J. and E. C. Traugott (1993). Grammaticalization. Cambridge: Cambridge University Press.

Hudson, R. (1990). English Word Grammar. Oxford: Blackwell Publishers Ltd.
Hudson, R. (1993). Heads in Grammatical Theory, Chapter Do we Have Heads in Our Minds?, pp. 266-291. Cambridge, UK: Cambridge University Press.

Iacobini, C. (2005). I verbi italiani come base di derivazione prefissale. In La formazione delle parole. Atti del XXXVII Congresso internazionale di studi della Societ di linguistica italiana (SLI). L'Aquila, 25-27 settembre 2003, Roma, pp. 289-307. Bulzoni.

Iacobini, C. and F. Masini (2007). Verb-particle constructions and prefixed verbs in italian: typology, diachrony and semantics. In G. Booij, B. Fradin, A. Ralli, and S. Scalise (Eds.), On-line Proceedings of the Fifth Mediterranean Morphology Meeting (MMM5) 15-18 settembre 2005), pp. 157-184. Università degli Studi di Bologna.

Ikegami, Y. (1987). Source vs. Goal : A Case of Linguistic Dissymmetry. In R. Dirven and G. Radden (Eds.), Concepts of Case, pp. 122-146. Tübingen: Günter Narr Verlag.

Imbert, C. (2008). Systems dynamics and functional motivations in Path coding. A typological description of Homeric Greek and Old English. Ph. D. thesis, CNRS / University of Lyon 2.

Jackendoff, R. (1983). Semantics and cognition. Cambridge, MA: MIT Press.
Jackendoff, R. (1990). Semantic structures. Cambridge, MA: MIT Press.
Jacobson, R. (1966). Linguistica e poetica. In L. Heilmann and L. Grassi (Eds.), Saggi di linguistica generale (Italian translation ed.)., pp. 181-218. Milano: Feltrinelli.

Jenset, G. B. (2008). Existential there beyond grammaticalization. In G. B. Jenset, C. M. D. Heggelund, Øystein, S. Wold, and A. Didriksen (Eds.), Linguistics in the Making. Selected papers from the Second Scandinavian PhD Conference in Linguistics and Philology in Bergen, June 4-6, 2007, pp. 57-75. Oslo: Novus Press.

Jenset, G. B. (2010). A Corpus-based Study on the Evolution of There: Statistical Analysis and Cognitive Interpretation. Ph. D. thesis, University of Bergen.

Katz, J. and J. Fodor (1964). The structure of a semantic theory. In J. Fodor and J. Katz (Eds.), The Structure of Language, pp. 479-518. Prentice Hall.

Kingsbury, P. and M. Palmer (2002). From Treebank to Propbank. In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas, Canary Islands.

Koch, U. (1993). The enhancement of a dependency parser for Latin. Technical Report AI-1993-03, Artificial Intelligence Programs, University of Georgia.

Korhonen, A. (2002). Subcategorization acquisition. Ph. D. thesis, University of Cambridge, Cambridge, UK.

Korhonen, A., Y. Krymolowski, and T. Briscoe (2006). A large subcategorization lexicon for Natural Language Processing Applications. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006). Genoa.

Koster, C. H. A. (1991). Affix grammars for natural languages. In Attribute Grammars, Applications and Systems, International Summer School SAGA. Prague, Volume 545 of Lecture Notes in Computer Science.

Koster, C. H. A. (2005). Constructing a parser for Latin. In Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science, Berlin/Heidelberg, pp. 48-59. Springer.

Kromann, M. T. (2003). The Danish Dependency Treebank and the underlying linguistic theory. In J. Nivre and E. Hinrichs (Eds.), Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT 2003). Växjö.

Kruengkrai, C., T. Charoenporn, V. Sornlertlamvanich, and H. Isahara (2004). Acquiring selectional preferences in a Thai lexical database. In Proceedings of First Joint Conference on Natural Language Processing (IJCNLP 2004). Sanya City, pp. 668-675.

Kullback, S. and R. A. Leibler (1951). On information and sufficiency. Annals of Mathematical Statistics 22(1), 79-86.

Kunze, C. (2000). Extension and Use of GermaNet, a Lexical-Semantic Database. In Proceedings of the 2nd International Conference on Language Resources and Evaluation. Athens, pp. 999-1002.

Kuryłowicz, J. (1975). The evolution of grammatical categories. In Esquisses linguistiques II, pp. 38-54. München: Fink.

Lakoff, G. (1987). Wome, Fire, and Dangerous Things: What categories reveal about the mind. Chicago: University of Chicago Press.

Lakusta, L. and B. Landau (2005). Starting at the end: the importance of goals in spatial language. Cognition 96, 1-33.

Langacker, R. W. (1977). Syntactic reanalysis. In C. N. Li (Ed.), Mechanisms of Syntactic Change, pp. 57-139. Austin: University of Texas Press.

Lapata, M. (1999). Acquiring lexical generalizations from corpora: a case study for diathesis alternations. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, pp. 397-404.

Lazzeroni, R. (1998). Divagazioni sulla degrammaticalizzazione. In G. Bernini, P. Cuzzolin, and P. Molinelli (Eds.), Ars linguistica. Studi offerti a Paolo Ramat, pp. 275283. Roma: Bulzoni.

Lehmann, C. (1983). Latin preverbs and cases. In H. Pinkster (Ed.), New studies in Latin Linguistics: selected papers from the International Colloquium on Latin Linguistics, 1981, Amsterdam/Philadelphia, pp. 144-165. John Benjamins.

Lenci, A., M. Johnson, and G. Lapesa (2010). Building an Italian FrameNet through Semi-automatic Corpus Analysis. In Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta. European Language Resources Association (ELRA).

Lenci, A., B. McGillivray, S. Montemagni, and V. Pirrelli (2008). Unsupervised acquisition of verb subcategorization frames from shallow-parsed corpora. In Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008). Marrakech, pp. 3000-3006. European Language Resources Association.

Lenci, A., S. Montemagni, and V. Pirrelli (2005). Testo e computer. Elementi di linguistica computazionale. Roma: Carocci.

Leonetti, M. (2004). Specificity and differential object marking in Spanish. Catalan Journal of Linguistics 3, 75-114.

Levin, B. (1993). English Verb Classes and Alternations. The University of Chicago Press.

Levin, B. and M. Rappaport Hovav (2005). Argument Realization. Cambridge: Cambridge University Press.

Li, H. and N. Abe (1995). Generalizing case frames using a thesaurus and the MDL principle. In Proceedings of Recent Advances in Natural Language Technology, pp. 239-248.

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

Lichtenberk, F. (1991). Approaches to grammaticalization, Volume I. Focus on theoretical and methodological issues, Chapter On the gradualness of grammaticalization. Amsterdam/Philadelphia: John Benjamins.

Light, M. and W. Greiff (2002). Statistical models for the induction and use of selectional preferences. Cognitive Science 26(3), 269-281.

Lightfoot, D. (2006). How new languages emerge. Cambridge: Cambridge University Press.

Lin, D. (1995). A Dependency-based Method for Evaluating Broad-Coverage Parsers. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI95). Montreal, Volume 2, pp. 1420-1425.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In Proceedings of COLING-ACL 1998, Montreal.

Longrée, D. (2005). Verbes préfixés et concurrences syntaxiques chez Tacite. In C. Moussy (Ed.), La composition et la préverbation en latin, pp. 293-307. Presses de l'Université Paris-Sorbonne.

Lopatková, M. (2003). Valency in the Prague Dependency Treebank: Building the Valency Lexicon. Prague Bulletin of Mathematical Linguistics 78-80, 37-60.

López Moreda, S. (1998). Interferencias semántico-sintácticas entre preverbios y preposiciones. In B. García-Hernández (Ed.), Estudios de Lingüística latina. Actas del IX Coloquio Internacional de Lingüística latina. Universidad Autonóma de Madrid, 1997, pp. 953-969. Ediciones Clásicas.

Lüdtke, J. (1996). Lexikon der Romanistischen Linguistik (LRL), Volume II, 1. Latein und Romanisch. Historisch-vergleichende Grammatik der romanischen Sprachen/Le latin et le roman. Grammaire historico-comparative des langues romanes, Chapter Gemeinromanische Tendenzen IV. Wortbildungslehre/Tendences communes aux langues romanes IV. Formation de mots, pp. 235-272. Tübingen: Max Niemeyer Verlag.

Luraghi, S. (1989). The relationship between prepositions and cases within Latin prepositional phrases. In G. Calboli (Ed.), Subordination and Other Topics in Latin. Proceedings of the Third Colloquium on Latin Linguistics, Bologna, 1-5 April 1985, Amsterdam/Philadelphia, pp. 253-271. John Benjamins.

Luraghi, S. (2003). On the meaning of prepositions and cases: the expression of semantic roles in ancient Greek. Amsterdam/Philadelphia: John Benjamins.

Luraghi, S. (2010b). Continuum Companion to Historical Linguistics, Chapter The rise (and possible downfall) of configurationality, pp. 212-229. London/New York: Continuum.

MacDonell, A. A. (1953). A Vedic grammar for students. Bombay/Calcutta/Madras: Oxford University.

Magerman, D. M. (1995). Statistical decision-tree models for parsing. In Proceedings of the 33rd Annual Meeting of the ACL. Boston, pp. 276-283.

Magni, E. (2008). The Evolution of Latin Word (Dis)order. In S. Scalise, E. Magni, and A. Bisetto (Eds.), Universals of Language Today, pp. 225-251. New York: Springer.

Marcus, M., B. Santorini, and M. A. Marcinkiewicz (1993). Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics 19(2), 313-330.

Masini, F. (2005). Multi-word expressions between syntax and the lexicon: the case of Italian verb-particle constructions. SKY journal of linguistics 18, 145-173.

MathWorks Inc., T. (2004). Matlab [7.0].
Matisoff, J. A. (1991). Approaches to Grammaticalization, Volume 2. Focus on types of grammatical markers, Chapter Areal and universal dimensions of grammaticalization on Lahu, pp. 383-453. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Mayrhofer, M. (1980). Zur Gestaltung des etymologischen Wörterbuchs einer "Großcorpus-Sprache". Wien: Österr. Akademie der Wissenschaften. Phil-Hist. Klasse.

McCarthy, D. (1997). Word Sense Disambiguation for acquisition of selectional preferences. In Proceedings of the ACL/EACL 97 Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications. Madrid, pp. 52-61.

McCarthy, D. (2001). Lexical acquisition at the syntax-semantics interface: diathesis alternations, subcategorization frames and selectional preferences. Ph. D. thesis, University of Sussex.

McDonald, R., F. Pereira, K. Ribarov, and J. Hajič (2005). Non-projective dependency parsing using spanning tree algorithms. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT 2005). Vancouver, Morristown, pp. 523-530. Association for Computational Linguistics.

McEnery, T. and A. Wilson (2001). Corpus linguistics. An introduction. Edinburgh: Edinburgh University Press.

McGillivray, B. and M. Passarotti (2009). The development of the Index Thomisticus Treebank Valency Lexicon. In L. Borin and P. Lendvai (Eds.), Proceedings of the Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH-SHELTEBR 2009). Athens, pp. 33-40.

McGillivray, B., M. Passarotti, and P. Ruffolo (2009). The Index Thomisticus Treebank Project: Annotation, Parsing and Valency Lexicon. TAL 50(2), 103-127.

McKeown, K. and V. Hatzivassiloglou (1993). Augmenting lexicons automatically: clustering semantically related adjectives. In Proceedings of the ARPA Workshop on Human Language Technology, pp. 272-277. Morgan Kaufman.

Meillet, A. (1921). Linguistique historique et linguistique générale, Chapter L'évolution des formes grammaticales, pp. 130-148. Paris: Champion.

Meillet, A. and J. Vendryes (1963). Traité de grammaire comparée des langues classiques. Paris: Librairie Ancienne Honoré Champion.

Meini, L. (2009). Dimensioni dello spazio nelle preposizioni. Pisa: Pisa University Press.

Meini, L. and B. McGillivray (2010). Between semantics and syntax: spatial verbs and prepositions in Latin. In Proceedings of the Space in Language Conference, Pisa October 8-10 2009.

Mellado Rodríguez, J. (2001). Determinación sintáctica y determinación semántica: diferencias y repercusión en algunas proposiciones latinas. In C. Moussy (Ed.), De lingua latina novae quaestiones. Actes du Xè Colloque International de Linguistique Latine. Paris-Sèvres, 1999, pp. 459-472. Peeters.

Mel'čuk, I. (1988). Dependency Syntax: Theory and Practice. Albany: State University of New York Press.

Mel'čuk, I. (2003). Dependenz und Valenz (Dependency and Valency), Volume 1: Ein Internationales Handbuch Zur Zeitgenossischen Forschung (An International Handbook of Contemporary Research), Chapter Levels of Dependency in Linguistic Description: Concepts and Problems. Mouton de Gruyter.

Mel'čuk, I. (2004a). Actants in semantics and syntax i: actants in semantics. Linguistics 42(1), 1-66.

Mel'čuk, I. (2004b). Actants in semantics and syntax ii: actants in syntax. Linguistics 42(1), 247-292.

Messiant, C., A. Korhonen, and T. Poibeau (2008). LexSchem: A large subcategorization lexicon for French verbs. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008). Marrakech.

Miller, G., C. Leacock, R. Tengi, and R. T. Bunker (1993). A semantic concordance. In Proceedings of the ARPA Workshop on Human Language Technology, pp. 303-308. Morgan Kaufman.

Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller (1990). Introduction to WordNet: an on-line lexical database. International Journal of Lexicography 3(4), 235-244.

Milroy, J. (1993). On the social origins of language change. In C. Jones (Ed.), Historical linguistics: Problems and perspectives, pp. 215-236. London: Longman.

Minozzi, S. (2009). The Latin WordNet Project. In P. Anreiter and M. Kienpointner (Eds.), Proceedings of the 15th International Colloquium on Latin Linguistics (ICLL), Innsbrucker Beitraege zur Sprachwissenschaft, Innsbruck.

Molinari, D. (2005). Les préverbés de facio. In C. Moussy (Ed.), La composition et la préverbation en latin, pp. 345-359. Presses de l'Université Paris-Sorbonne.

Mosteller, F. (1968). Association and estimation in contingency tables. Journal of the American Statistical Association 321(63), 1-28.

Moussy, C. (2005). La polysémie du préverbe com-. In C. Moussy (Ed.), La composition et la préverbation en latin, pp. 243-262. Presses de l'Université Paris-Sorbonne.

Navigli, R., P. Velardi, A. Cucchiarelli, and F. Neri (2004). Extending and Enriching WordNet with OntoLearn. In Proceedings of the Second Global WordNet Conference, Brno, Czech Republic, pp. 279-284.

Nivre, J. (2005a). Dependency Grammar and Dependency Parsing. Technical report, Växjö University: School of Mathematics and Systems Engineering.

Nivre, J. and M. Scholz (2004). Deterministic dependency parsing of English text. In Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004). Geneva, pp. 64.
Oniga, R. (2005). Composition et préverbation en latin: problèmes de typologie. In C. Moussy (Ed.), La composition et la préverbation en latin, pp. 211-227. Presses de l'Université Paris-Sorbonne.

Oniga, R. (2007). Il latino. Breve introduzione linguistica. Franco Angeli.

Palmer, M., D. Gildea, and P. Kingsbury (2005, March). The Proposition Bank: An Annotated Corpus of Semantic Roles. Computational Linguistics 31(1), 71-106.

Pantel, P., R. Bhagat, B. Coppola, T. Chklovski, and E. Hovy (2007). ISP: Learning inferential selectional preferences. In Proceedings of North American Association for Computational Linguistics / Human Language Technology (NAACL HLT 2007), Rochester, pp. 564-571.

Passarotti, M. (2007a). LEMLAT. Uno strumento per la lemmatizzazione morfologica automatica del latino. In F. Citti and T. Del Vecchio (Eds.), From Manuscript to Digital Text. Problems of Interpretation and Markup. Proceedings of the Colloquium (Bologna, June 12th 2003). Roma, pp. 107-128.

Passarotti, M. (2007b). Verso il Lessico Tomistico Biculturale. La treebank dell'Index Thomisticus. In R. Petrilli and D. Femia (Eds.), Il filo del discorso. Intrecci testuali, articolazioni linguistiche, composizioni logiche. Atti del XIII Congresso Nazionale della Società di Filosofia del Linguaggio. Viterbo, pp. 187-205.

Passarotti, M. (2010). Leaving Behind the Less-Resourced Status. The Case of Latin through the Experience of the Index Thomisticus Treebank. In Proceedings of the 7th SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for LessResourced Languages, LREC 2010, La Valletta, Malta, pp. 27-32.

Passarotti, M. and F. Dell'Orletta (2010). Improvements in Parsing the Index Thomisticus Treebank. Revision, Combination and a Feature Model for Medieval Latin. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010). La Valletta, Malta, pp. 1964-1971. European Language Resources Association (ELRA).

Passarotti, M. and P. Ruffolo (2004). L'utilizzo del lemmatizzatore LEMLAT per una sistematizzazione dell'omografia in latino. Evphrosyne. Revista de Filologia Clásica XXXII, 99-110.

Passarotti, M. and P. Ruffolo (2009). Parsing the Index Thomisticus Treebank. Some preliminary results. In P. Anreiter and M. Kienpointner (Eds.), Proceedings of the 15th International Colloquium on Latin Linguistics, Innsbrucker Beiträge zur Sprachwissenschaft, Innsbruck.

Pearl, J. (1988). Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufman.

Pereira, F., N. Tishby, and L. Lee (1993). Distributional clustering of English words. In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, pp. 183-190.

Peterson, D. (2007). Applicative Constructions. Oxford: Oxford University Press.
Pinkster, H. (1990). Latin Syntax and Semantics. London: Routledge.
Pinkster, H. (2010). The use of the dative with Latin compounds. STUF 63, 32-43.
Poccetti, P., D. Poli, and C. Santini (1999). Una storia della lingua latina. Roma: Carocci.

R Development Core Team (2008). R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Ramat, P. (1992). Thought on degrammaticalization. Linguistics 30, 549-560.
Rasinger, S. M. (2008). Quantitative research in linguistics. London/New York: Continuum.

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

Resnik, P. (1996). Selectional constraints: an information-theoretic model and its computational realization. Cognition 61, 127-199.

Resnik, P. (1997). Selectional preference and sense disambiguation. In Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why What and How?, Washington, pp. 79-86.

Ribas, F. (1994). An experiment on learning appropriate selection restrictions from a parsed corpus. In Proceedings of the International Conference on Computational Linguistics, COLING-94. Kyoto, Japan, pp. 769-774.

Ribas, F. (1995). On Acquiring Appropriate Selectional Restrictions from Corpora Using a Semantic Taxonomy. Ph. D. thesis, University of Catalonia.

Rissanen, J. (1989). Stochastic complexity in statistical inquiry. Singapore: World Scientific Publishing Co.

Rohlfs, G. (1970). Grammatica storica della lingua italiana e dei suoi dialetti. Giulio Einaudi.

Romagno, D. (2003). Azionalità e transitività: il caso dei preverbi latini. Archivio Glottologico Italiano 88, 156-170.

Romagno, D. (2004). Ancora su preverbazione e sistemi verbali: il caso dei preverbi greci. Archivio Glottologico Italiano 89, 165-180.

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

Rosén, H. (1999). Latine loqui. Trends and directions in the crystallization of Classical Latin. München: Fink.

Rousseau, A. (Ed.) (1995). Les Préverbes dans les langues d'Europe: introduction à l'étude de la préverbation. Lille: Presses Universitaires du Septentrion.

Rovai, F. (2008). Tratti attivi in latino. Il caso del genere. Ph. D. thesis, Università di Pisa, Pisa.

Ruppenhofer, J., M. Ellsworth, M. R. L. Petruck, C. R. Johnson, and J. Scheffczyk (2006). FrameNet II: Extended Theory and Practice. Berkeley, California: International Computer Science Institute.

Salvi, G. and L. Vanelli (1992). Grammatica essenziale di riferimento della lingua italiana. Istituto Geografico De Agostini. Le Monnier.

Saussure, F. d. (1968). Corso di linguistica generale. Laterza.
Schmid, G. (1994). TreeTagger - a language independent part-of-speech tagger. Available at http://www.ims.uni-stuttgart.de/Tools/DecisionTreeTagger.html.

Schulte im Walde, S. (2003). Experiments on the automatic induction of German semantic verb classes. Ph. D. thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Schulte im Walde, S. (2004). GermaNet synsets as selectional preferences in semantic verb clustering. Journal for Computational Linguistics and Language Technology 19(1/2), 69-79.

Schwarze, C. (1985). "Uscire" e "andare fuori": struttura sintattica e semantica lessicale. In A. Franchi de Bellis and L. M. Savoia (Eds.), Sintassi e morfologia della lingua italiana d'uso. Teorie e applicazioni descrittive. Atti del XXIV Congresso Internazionale SLI., Roma, pp. 355-371. Bulzoni.

Serianni, L. (1989). Grammatica italiana. Italiano comune e lingua letteraria. Torino: UTET Libreria.

Sgall, P., E. Hajicová, and J. Panevová (1986). The Meaning of the Sentence in its Semantic and Pragmatic Aspects. D. Reidel, Dordrecht, NL.

Simone, R. (1996). Esistono verbi sintagmatici in italiano? Cuadernos de Filologia Italiana 3, 47-61.

Simone, R. (1997). Lessico e grammatica. Teorie linguistiche e applicazioni lessicografiche, Chapter Esistono verbi sintagmatici in italiano?, pp. 155-170. Roma: Bulzoni.

Simone, R. (2008). I verbi sintagmatici come costruzione e come categoria. In M. Cini (Ed.), I verbi sintagmatici in italiano e nelle varietà dialettali. Stato dell'arte e prospettive di ricerca. Atti delle giornate di studio Torino, 19-20 febbraio 2007, Frankfurt am Main. Peter Lang.

Sinha, C. and T. Kuteva (1995). Distributed spatial semantics. Nordic Journal of Linguistics 18, 167-199.

Slobin, D. I. (2004). The many ways to search for a frog: linguistic typology and the expression of motion events. In S. Strömqvist and L. Verhoeven (Eds.), Relating events in narrative, Volume 2. Typological and contextual perspective, pp. 219-257. Mahwah, NJ: Lawrence Erlbaum Associates.

Smadja, F. (1993). Retrieving collocations from text: Xtract. Computational Linguistics. Special Issue on Using Large Corpora 19(1), 143-177.

Somers, H. (1984). On the validity of the complement-adjunct distinction in valency grammar. Linguistics 22, 507-530.

Spreafico, L. (2009). Problemi di tipologia lessicale. I verbi di moto nello Standard Average European. Bulzoni.

Stefanowitsch, A. and A. Rohde (2004). The goal bias in the encoding of motion events. In K.-U. Panther and G. Radden (Eds.), Motivation in Grammar, pp. 249268. Berlin/New York: Mouton de Gruyter.

Stolova, N. I. (2008). From satellite-framed Latin to verb-framed Romance: Late Latin as an intermediate stage. In R. Wright (Ed.), Latin vulgaire-latin tardif VIII: actes du VIIIe Colloque international sur le latin vulgaire et tardif, Oxford, 6-9 septembre 2006, Hildesheim/New York, pp. 253-262. Olms-Weidmann.

Sweetser, E. (1988). Grammaticalization and semantic bleaching. In S. Axmaker, A. Jaisser, and H. Singmaster (Eds.), Grammaticalization, Proceedings of the fourteenth annual meeting of the Berkeley Linguistics Society, 13-15 February 1988, Berkeley, pp. 389-405. Berkeley Linguistic Society.

Talmy, L. (1991). Path to realization: a typology of event conflation. In Proceedings of the 17th Annaul Meeting of the Berkeley Linguistics Society, Volume 17, Berkeley, CA, pp. 480-591. Berkeley Linguistic Society.

Talmy, L. (2000). Toward a cognitive semantics, Volume II. Bradford Books.
Tekavčić, P. (1972). Grammatica storica dell'italiano, Volume II Morfosintassi; III Lessico. Bologna: Il Mulino.

Tesnière, L. (1959). Éléments de syntaxe structurale. Paris: Klincksieck.
Tognini-Bonelli, E. (2001). Corpus Linguistics at work. Amsterdam/Philadelphia: John Benjamins.

Touratier, C. (1994). Syntaxe latine. Louvain-La-Neuve: Peeters.
Touratier, C. (2001). Rection et valence. In C. Moussy and J. Dangel (Eds.), De lingua latina novae quaestiones. Actes du dixième Colloque International de Linguistique Latine. Paris-Sèvres, 19-23 avril 1999, pp. 641-649.

Traina, A. (1955). Esegesi e sintassi. Studi di sintassi latina, Chapter L'apposizione epitetica dei nomi di città nei determinativi di luogo, pp. 1-25. Padova: Liviana Editrice.

Traina, A. (1999). Forma e suono da Plauto a Pascoli. Bologna: Pàtron.
Traina, A. and T. Bertotti (1993). Sintassi normativa della lingua latina. Cappelli.
Traugott, E. C. (2006). The Handbook of Historical Linguistics, Chapter Constructions in grammaticalization, pp. 626-647. Blackwell.

Tukey, J. W. (1977). Exploratory Data Analysis. Reading, MA: Addison-Wesley.
Untermann, J. (1983). Indogermanische Restsprachen als Gegenstand der Indogermanistik. In E. Vineis (Ed.), Le lingue indoeuropee di frammentaria attestazione. Die indogermanischen Restsprachen. Atti del convegno della Societa italiana di glottologia e della Indogermanische Gesellschaft. Udine, 22-24 settembre 1981, Pisa, pp. 11-28. Società Italiana di Glottologia: Giardini.

Urešová, Z. (2004). The Verbal Valency in the Prague Dependency Treebank from the Annotator's Point of View. Bratislava: Jazykovedný ústav L. Štúra, SAV.
Urso, A. M. (1999). I preverbi nel latino tardo: il caso di Celio Aureliano. In R. Wright (Ed.), Latin vulgaire-latin tardif. Actes du VIII Colloque international sur le latin vulgaire et tardif, Oxford, 6-9 septembre 2006, Hildesheim-Zrich-New York, pp. 292300. Olms-Weidmann.

Ushioda, A., D. A. Evans, T. Gibson, and A. Waibel (1993). The automatic acquisition of frequencies of verb subcategorization frames from tagged corpora. In B. Boguraev and J. Pustejovsky (Eds.), Proceedings of the Workshop on Acquisition of Lexical Knowledge from Text.

Väänänen, V. (1971). Introduzione al latino volgare. Bologna: Pàtron.
Van der Beek, L., G. Bouma, R. Malouf, and G. van Noord (2002). The Alpino Dependency Treebank. In M. Theune, A. Nijholt, and H. Hondorp (Eds.), Proceedings of the Twelfth Meeting of Computational Linguistics in the Netherlands (CLIN 2001). Rodopi, Amsterdam, pp. 8-22.

Van Laer, S. (1998). Le préverbe ob- en latin: considérations sémantiques. In B. GarcíaHernández (Ed.), Estudios de Lingüística latina. Actas del IX Coloquio Internacional de Lingüística latina. Universidad Autonóma de Madrid, 1997, pp. 1043-1059. Ediciones Clásicas.

Van Laer, S. (2005). Per et les procès gradables. In C. Moussy (Ed.), La composition et la préverbation en latin, pp. 321-343. Presses de l'Université Paris-Sorbonne.

Vicario, F. (1997). I verbi analitici in friulano. Milano: Franco Angeli.
Vincent, N. (1999). The evolution of c-structure: prepositions and PPs from IndoEuropean to Romance. Linguistics 37, 1111-1154.

Viti, C. (2008). From space words to transitive markers: the case of ancient Greek en 'in'. Transactions of the Philological Society 106(3), 375-413.
von Eye, A. (1990). Introduction to configural frequency analysis. The search for types and antitypes in cross-classification. Cambridge: Cambridge University Press.

Wackernagel, J. (1892). Über ein Gesetz der indogermanischen Wortstellung. Indogermanische Forschungen 1, 333-436.

Wälchli, B. (2000). A typology of displacement (with special reference to Latvian). Sprachtypologie und Universalienforschung 54(3), 298-323.

Whitney, W. D. (1989). Sanskrit grammar. Delhi: Motilal Banarsidass.
Wilks, Y. (1986). An intelligent analyzer and understander of English. In S.-J. K. Grosz, B. and B. Webber (Eds.), Readings in Natural Language Processing., pp. 264-274. California: Morgan Kaufmann.

Wilks, Y. and M. Stevenson (1998). Optimising combinations of knowledge sources for word sense disambiguation. In Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics, Montreal, Canada, pp. 1398-1402.

Xia, F., P.-M. (2001). Converting dependency structures to phrase structures. In Proceedings of the First International Conference on Human Language Technology Research. San Diego, pp. 1-5.

Yamada, H., M.-Y. (2003). Statistical dependency analysis with support vector machines. In Proceedings of the 8th International Workshop on Parsing Technologies (IWPT). Nancy, pp. 195-206.

Zeman, D. and A. Sarkar (2000b). Learning verb subcategorization from corpora: Counting frame subsets. In Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC2000), Athens, Greece.

Zwicky, A. M. (1985). Heads. Journal of Linguistics 21, 1-29.

## Index

Abney and Light (1999), vi, 85, 86, 199, Brinton and Traugott (2005), 126, 266 $264 \quad$ Brinton (1988), 127, 266
Agirre et al. (2000), 100, 264
Agirre and Martinez (2001), vi, 85, 86, 200, 264

Briscoe and Carroll (1997), 23, 49, 50, 52, 266
Buchholz and Marsi (2006), 41, 266
Alishahi (2008), 1, 5, 81, 91, 96, 99, 102, Busa (1980), 3, 37, 38, 266
104, 107, 111, 117, 166, 215, 217, Buysschaert (1982), 16, 266
264
Allen (1981), 126, 127, 264
Alonso et al. (2007), 26, 51, 264
Apollon (2003), 140, 264
Aquinas (1858), 263
Attardi (2006), 41, 264
Böhmová et al. (2003), 21, 47, 265
Bühler (1934), 43, 266
Baayen (2008), 203, 264
Baker et al. (1998), 26, 48, 208, 264
Baldi (1979), 124, 264
Bamman et al. (2007a), 59, 264
Bamman et al. (2007b), 54, 264
Bamman et al. (2008), 44, 97, 265
Bamman and Crane (2006), 32, 264
Bamman and Crane (2007), 1, 40, 53, 264
Bamman and Crane (2008), vi, 11, 41, 53, 78, 79, 91, 113, 208, 264
Basili et al. (1997), 51, 265
Bentivogli and Pianta (2004), 100, 265
Benveniste (1949), 120, 121, 265
Benzécri (1973), 154, 155, 244, 265
Bianco (1996), 45-47, 263
Bilisoly (2008), 203, 265
Bloomfield (1933), 10, 265
Blumenthal and Rovere (1998), 45, 46, 263
Booij and van Kemenade (2003), 119, 121, 265
Bortolussi (2005), 121, 132, 135, 143, 265
Bossong (1983), 162, 265
Brachet (2000), 144, 161, 265
Brachet (2005), 121, 123, 265
Brent and Berwick (1991), 23, 50, 52, 265
Brent (1991), 49, 265
Caesar (1914), 263
Calboli (2005), 121, 266
Calonghi (1967), 34, 263
Calvo and Gelbukh (2004), 85, 87, 111, 266
Campbell (2001), 124, 266
Carnie (2007), 25, 266
Carroll et al. (1998), 44, 55, 79, 266
Carvalho (1983), 30, 266
Carvalho (1996), 30, 266
Carvalho (1998), 121, 267
Castellón et al. (2006), 51, 267
Castiglioni and Mariotti (1993), 100, 267
Chomsky (1965), 24, 267
Chomsky (1970), 24, 267
Chomsky (1981), 25, 267
Ciaramita and Johnson (2000), vi, 85, 86, 199, 267
Cicero (1909), 263
Clark and Weir (1999), vi, 85, 86, 199, 267
Cohen (1988), 149, 233, 267
Coleman (1991), 134-136, 267
Collins (1996), 55, 267
Collins (1997), 23, 267
Covington (1990), 41, 267
Craig and Hale (1988), 121, 267
Crane (1991), 39, 267
Crispi (1919), 263
Crocco Galèas and Iacobini (1992), 126, 268
Croft (2000), 10, 268
Cuzzolin et al. (2006), 120, 122, 124, 125, 268
Cuzzolin (1995), 120, 123-126, 268
Dardano (1978), 127, 268

Dell'Orletta et al. (2007), 98, 268
Dempster and Rubin (1977), 91, 208, 268
Denison (2003), 166, 268
De Vivo (1994), 162, 268
Dikken (1995), 129, 268
Dowty (1991), 17, 268
Dufresne et al. (2003), 126, 127, 268
Dunning (1993), 197, 268
Durante (1981), 127, 268
Echarte Cossío (1998), 120, 121, 268
Engel and Savin (1983), 46, 263
Erk (2007), vi, 91, 111, 118, 208, 268
Ernout and Thomas (1972), 131, 134, 138, 268
Evert and Baroni (2006), 73, 268
Evert (2006), 11, 268
Faraway (2006), 240, 269
Fillmore et al. (2003), 26, 125, 269
Fillmore (1968), 17, 269
Fillmore (1985), 26, 269
Fillmore (1995), 23, 269
Fillmore (2007), 48, 269
Firth (1957), 87, 269
Fischer (1971), 162, 269
Fischer (2007), 126, 269
Forcellini (1940), 32, 263
Francis (1979), 200, 269
Fraser (1990), 22, 269
Fruyt (1998), 120, 123, 124, 269
Götz-Votteler (2007), 21, 45, 270
Gaide (2005), 121, 269
Gale and Sampson (1995), 107, 269
García-Hernández (2004), 127, 269
García-Hernández (2005), 121, 142, 269
Garcìa Jurado (1991), 127, 269
Geeraerts (2006), 168, 270
Gelman et al. (2004a), 193, 270
Gelman et al. (2009), 240, 270
Giacomelli (1993), 9, 270
Glare (Glare), 263
Goggi (1997), 51, 270
Goldberg (1995), 26, 27, 81, 125, 270
Goldberg (2006), 18, 270
Gorard (2003), 7, 11, 232, 234, 235, 270
Greenacre and Nenadic (2009), 251, 270
Greenacre (2007), 244, 270
Greenough (1882), 263
Griffe (1995), 29, 270
Gropen et al. (1989), 25, 270
Gruber (1976), 17, 270

Haegeman (1996), 25, 270
Hajič et al. (1999), 40, 58, 270
Hajič et al. (2003), 22, 47, 270
Halácsy et al. (2007), 41, 271
Hamp and Feldweg (1997), 87, 271
Happ (1976a), 29, 45, 78, 184, 271
Happ (1976b), 45, 271
Harabagiu and Moldovan (2000), 100, 271
Harrell et al. (1996), 109, 271
Haug and Jøndal (2008), 79, 271
Haug (2009), 122, 271
Haverling (2000), 126, 271
Heine et al. (1991), 121, 124, 125, 271
Heine and Kuteva (2002), 126, 271
Helbig (1992), 16, 21, 45, 271
Hellwig (1986), 19, 271
Herbst et al. (1980), 16, 271
Herbst et al. (2004), 28, 45, 47, 263
Heslin (1987), 121, 271
Hindle (1990), 208, 271
Hinrichs and Telljohann (2009), 191, 271
Hinton (2004), 225, 231-235, 272
Hlaváčková and Kadlec (2006), 79, 272
Hofmann (1980), 162, 272
Hopper and Traugott (1993), 125, 272
Hudson (1990), 19, 20, 272
Hudson (1993), 20, 272
Iacobini and Masini (2007), 127, 272
Iacobini (2005), 120, 127, 272
Ikegami (1987), 137, 159, 272
Imbert (2008), 120, 121, 127, 135-137, 159, 163, 272
Jackendoff (1983), 25, 272
Jackendoff (1990), 130, 272
Jacobson (1966), 164, 272
Jenset (2008), 126, 272
Jenset (2010), 12, 272
Jerome (Jerome), 263
Katz and Fodor (1964), 17, 272
Kingsbury and Palmer (2002), 48, 272
Koch (1993), 41, 273
Korhonen et al. (2006), 52, 273
Korhonen (2002), 18, 22, 26, 49-52, 273
Koster (1991), 41, 273
Koster (2005), 41, 273
Kromann (2003), 55, 273
Kruengkrai et al. (2004), 85, 273
Kullback and Leibler (1951), 85, 196, 273
Kunze (2000), 87, 273
Kuryłowicz (1975), 124, 273

Lüdtke (1996), 127, 275
López Moreda (1998), 120, 126, 134, 136, Meini and McGillivray (2010), 122, 145, 159, 275
Lakoff (1987), 26, 273
Lakusta and Landau (2005), 137, 159, 273
Langacker (1977), 123, 273
Lapata (1999), 49, 273
Lazzeroni (1998), 126, 273
Lehmann (1983), 30, 120, 121, 131, 132, 134, 274
Lenci et al. (2005), 7, 274
Lenci et al. (2008), 5, 23, 26, 51, 52, 274
Lenci et al. (2010), 52, 274
Leonetti (2004), 163, 274
Levin and Rappaport Hovav (2005), 17, 18, 26, 274
Levin (1993), 5, 18, 25, 27, 51, 114, 274
Lewis and Short (1879), 33, 53, 79, 100, 106, 142, 143, 171, 263
Li and Abe (1995), 49, 86, 197, 198, 274
Li and Abe (1998), vi, xix, 85, 86, 197, 198, 274
Lichtenberk (1991), 125, 274
Light and Greiff (2002), 82, 199, 274
Lightfoot (2006), 166, 274
Lin (1995), 23, 55, 274
Lin (1998), 208, 274
Longrée (2005), 121, 135, 138, 153, 159, 164, 274
Lopatková (2003), 47, 275
Luraghi (1989), 123, 275
Luraghi (2003), 122, 275
Luraghi (2010b), 120, 122-124, 275
MacDonell (1953), 122, 275
Magerman (1995), 23, 275
Magni (2008), 124, 275
Marcus et al. (1993), 5, 40, 50, 85, 275
Masini (2005), 120, 127, 129, 275
Matisoff (1991), 126, 275
Mayrhofer (1980), 9, 275
McCarthy (1997), 111, 276
McCarthy (2001), vi, 18, 44, 49, 82, 85, 86, 114, 198, 207, 276
McDonald et al. (2005), 113, 276
McEnery and Wilson (2001), 7, 276
McGillivray et al. (2009), 5, 41, 113, 276
McGillivray and Passarotti (2009), 53, 276
McKeown and Hatzivassiloglou (1993), 90, 276
Meillet and Vendryes (1963), 122, 126, 276

Meillet (1921), 124, 276 276
Meini (2009), 121, 276
Mel'čuk (1988), 18-20, 276
Mel'čuk (2003), 19, 23, 276
Mel'čuk (2004a), 21, 277
Mel'čuk (2004b), 21, 277
Mellado Rodríguez (2001), 120, 276
Messiant et al. (2008), 52, 277
Miller et al. (1990), 5, 42, 277
Miller et al. (1993), 86, 200, 277
Milroy (1993), 166, 277
Minozzi (2009), 1, 5, 42, 81, 277
Molinari (2005), 121, 277
Mosteller (1968), 233, 277
Moussy (2005), 121, 277
Navigli et al. (2004), 100, 277
Nivre and Scholz (2004), 113, 277
Nivre (2005a), 19, 277
Oniga (2005), 120, 123, 124, 131, 132, 138, 277
Oniga (2007), 29, 277
Ovid (1892), 263
Palmer et al. (2005), 48, 277
Pantel et al. (2007), 85, 87, 278
Passarotti and Ruffolo (2004), 99, 143, 278
Passarotti and Ruffolo (2009), 11, 41, 79, 278
Passarotti and Dell'Orletta (2010), 41, 118, 278
Passarotti (2007a), 39, 278
Passarotti (2007b), 1, 40, 53, 278
Passarotti (2010), 5, 278
Pearl (1988), 199, 278
Pereira et al. (1993), 90, 278
Peterson (2007), 135, 278
Petronius (1913), 263
Pinkster (1990), 18, 29-31, 58, 278
Pinkster (2010), 135, 278
Plautus (1895), 263
Poccetti et al. (1999), 9, 10, 159, 278
Propertius (1995), 263
Ramat (1992), 126, 279
Rasinger (2008), 11, 140, 279
Resnik (1993), vi, 49, 84, 85, 195, 279
Resnik (1996), 111, 279
Resnik (1997), 84, 85, 279
Ribas (1994), 49, 82, 279
Ribas (1995), vi, 84, 85, 197, 279

Rissanen (1989), 86, 279
Rohlfs (1970), 127, 129, 135, 279
Romagno (2003), 131, 279
Romagno (2004), 121, 279
Rooth et al. (1999a), vi, 91, 207, 279
Rosén (1999), 9, 10, 135, 136, 159, 279
Rousseau (1995), 121, 279
Rovai (2008), 30, 279
Ruppenhofer et al. (2006), 48, 279
Salvi and Vanelli (1992), 126, 279
Saussure (1968), 122, 279
Schmid (1994), 41, 280
Schulte im Walde (2003), 5, 26, 44, 90, 114, 280
Schulte im Walde (2004), 85, 87, 280
Schwarze (1985), 130, 280
Serianni (1989), 127, 280
Sgall et al. (1986), 19, 21, 47, 280
Simone (1996), 130, 280
Simone (1997), 127, 129, 280
Simone (2008), 120, 129, 280
Sinha and Kuteva (1995), 128, 137, 280
Slobin (2004), 129, 280
Smadja (1993), 49, 280
Somers (1984), 16, 280
Spreafico (2009), 121, 280
Stefanowitsch and Rohde (2004), 159, 280
Stolova (2008), 130, 280
Sweetser (1988), 135, 281
Talmy (1991), 128, 281
Talmy (2000), 121, 127-129, 131, 281
Tekavčić (1972), 126, 127, 281
Tesnière (1959), 18-20, 43, 281
Tognini-Bonelli (2001), 3, 77, 281
Touratier (1994), 127, 281
Touratier (2001), 121, 281
Traina and Bertotti (1993), 134, 281
Traina (1955), 134, 281
Traina (1999), 162, 281
Traugott (2006), 124, 125, 281
Tukey (1977), 140, 281
Untermann (1983), 9, 281
Urešová (2004), 44, 281
Urso (1999), 127, 281
Ushioda et al. (1993), 23, 49, 50, 52, 281
Väänänen (1971), 10, 159, 282
Van Laer (1998), 121, 282
Van Laer (2005), 121, 282
Van der Beek et al. (2002), 55, 282
Vicario (1997), 126, 127, 282

Vincent (1999), 126, 135, 282
Viti (2008), 122, 135, 138, 144, 282
Wälchli (2000), 129, 282
Wackernagel (1892), 123, 282
Whitney (1989), 122, 282
Wilks and Stevenson (1998), 198, 282
Wilks (1986), 17, 282
Xia (2001), 23, 282
Yamada (2003), 23, 282
Zeman and Sarkar (2000b), 51, 282
Zwicky (1985), 20, 283
tll (2002), 263
von Eye (1990), 163, 259, 282
MathWorks Inc. (2004), 207, 275
R Development Core Team (2008), 140, 188, 223, 225, 236, 251, 278
adjunct, 15-17, 19-21, 24, 27-29, 32-35, $45,46,50,51,65,77,79,136$, 176
annotation, $19,21,22,25,30,32,86,92$, 97, 172, 174, 176
argument, $1,5,13,15-35,37,43-47,49-$ 54, 59, 65, 67-69, 74-81, 84, 85, 91-93, 96-98, 102, 104, 108, 111, $117,119,122,125,130,131,136$, $139,141,143,144,152,153,159$, 164-167, 173-177, 179-185, 187, 196, 197, 199, 200, 208, 209, 216, 225
filler, $82,85,91,93,98,104,111,113$, 114
position, 82, 85-87, 92, 93, 98, 107109, 195, 216, 217
realisation, 29, 30, 43, 69
slot, 111
spatial, 119, 134, 144, 160
structure, $1,5,12,13,18,19,25-27$, $29,30,32,34,35,37,43-45,49$, 54, 77, 82, 92, 97, 99, 119, 121, $122,126,127,130,131,139,141$, 144, 166, 180, 182

Chi-square test, $57,76,146,150,152,160$, 231-233, 243, 259, 260
Configural Frequency Analysis, 258-260
Configurational Frequency Analysis, 259
corpus, $3,4,6,11,13,26,28,29,33-35$, 166, 195-198, 200, 201, 208, 211, 214
-based, 1, 3, 4, 6, 7, 9, 13, 126, 138, $139,141,144,163,165,166$
-driven, 141
analysis, 121, 231
Correspondence Analysis, 243, 244, 248251
Multiple, 243, 251-258, 260
Dependency, 19-23, 27, 29, 40, 43, 47, 51, $53,55-58,60-63,66,67,171,172$, 174-178, 181
annotation, 19, 22
Grammar, 18-23, 25, 27, 29, 35, 40, 92
parser, 41
Generative
Grammar, 23-25, 29
grammaticalization, 120-122, 124-126, 134, $137,159,163$
lexicalization, 120-122, 125-127, 129, 130, 133, 136-138, 159, 163
linear model, 152, 231, 235, 238
generalized, 234
mixed effect model, 152, 234, 239, 242
parsing, 4, 5, 11, 22, 113
Phrase Structure Grammar, 35
preverb, $1,4,6,7,12,13,30,55,70,76$, 119-127, 130-139, 142-145, 152154, 157-160, 162-164, 166, 231, 234, 238-240, 242, 243, 250, 251, 253, 255-257
satellite, 121, 137, 159
spatial, $15,130,131,134,139,142$, 143, 167
preverbed verb, 119-122, 124-127, 130-146, $152-155,160,161,163,164$
satellite, 128-130, 133, 159
-framed, 129, 130, 167
selectional
association, 196
score, 196, 197, 208
preferences, $1,6,7,12,13,15,17,18$, $32-35,49,81,83,85-87,91,92$, $94,96,99,103,107-109,117,153$, $165,167,196$
acquisition of, $6,7,12,13,42,44$, $69,81-85,87,91,92,100,109$, 111, 114, 166, 167
subcategorization, $12,15,18,22-27,32$, $35,43-45,49,50,52,171,176$
acquisition of, 23, 43, 49-53
class, 50, 54, 55, 68, 69, 74, 174, 181, 182
extraction of, $43,50,53,55,77,79$
frame, $50-54,68,69,74,76,79,167$, 181, 182, 184
lexicon, $1,6,7,12,13,15,18,19,22$, $25,28,29,32,42-45,47,49,50$, 81, 82, 86, 91, 96, 111, 114, 117, $122,139,141,143,144,165,167$, $173,174,180,181$
pattern, $43,46,47,49-51,68,165$
structure, $23,54,68,72,75,76,96$, $114,115,165,166,171,173,181$, 184, 188
treebank, 1, 5-7, 12, 13, 19, 22, 30, 40, 81, 85, 91, 96-99, 113, 141, 143, 144, $153,166,167,216$
valency, 19, 27, 32, 43, 68, 79, 98, 131, 132, $166,174,183$
lexicon, 19, 21, 22, 31, 32, 35, 47
pattern, 22
verb-framed, 129, 130, 167
WordNet, $1,5,6,13,42,82-87,92-94,96$, $98,100,102,107,109,114,116$, 118, 167
Latin, $1,5,6,13,42,81,93-100,105$, 108-110, 114, 117, 118, 166, 193, 209, 218
Multi-, 100-102, 105, 209
Princeton, 98-100, 209


[^0]:    ${ }^{1}$ See, for example, the website of the last edition of the LaTeCH workshop in 2010: http://ilk.uvt.nl/LaTeCH2010/.

[^1]:    ${ }^{2}$ For a discussion of corpora and their representativeness see section 1.3 .

[^2]:    ${ }^{1}$ English translation (Nivre, 2005a): "[1.2] The sentence is an organized whole, the constituent elements of which are words. [1.3] Every word that belongs to a sentence ceases by itself to be isolated as in the dictionary. Between the word and its neighbors, the mind perceives connections, the totality of which forms the structure of the sentence. [...] [2.1]. The structural connections establish dependency relations between the words. Each connection in principle unites a superior term and an inferior term.

[^3]:    [2.2] The superior term receives the name governor. The inferior term receives the name subordinate. Thus, in the sentence Alfred parle [...], parle is the governor and Alfred the subordinate."
    ${ }^{2}$ (Our translation) 'The verbal node expresses a whole little drama. As a drama, it implies a process and, most of the time, actors and circumstances. Transposed from the dramatic reality to structural syntax, the process, the actors and the circumstances respectively become the verb, the actants and the circumstants".

[^4]:    ${ }^{3}$ Despite these differences, DG and PSG structures can be automatically converted into each other (Magerman, 1995; Lin, 1995; Collins, 1997; Xia, 2001; Yamada, 2003). On the one hand, a phrase can be extracted from a DG tree in the form of a word plus its dependents. On the other hand, the dependency relations are expressed in a PSG tree through the embedding of the constituents; this is possible only after specific rules have defined which words are the heads of the phrase structures. In spite of these formal similarities, there are still theoretical differences that set the two frameworks apart in terms of views on language and goals of linguistic analysis (cfr. Fillmore 1995), and a full account of these differences falls outside the scope of this thesis.

[^5]:    ${ }^{4}$ An example of a two-valency noun is laudator 'praiser' ( $X_{\text {nom }}$ laudator $Y_{\text {gen }}$ est ' X is Y 's praiser'), an example of a two-valency adjective is similis 'similar' ( $X_{n o m}$ similis $Y_{\text {gen }}$ est ' X is similar to Y ').

[^6]:    ${ }^{5} \mathrm{http}: / /$ www.lib.uchicago.edu/efts/PERSEUS/Reference/lewisandshort.html.
    ${ }^{6}$ We will here refer to the CD-ROM edition published in 2002, based on the paper edition first published in 1900 in Leipzig.

[^7]:    ${ }^{1}$ http://www.mqdq.it/mqdq/home.jsp
    ${ }^{2}$ http://www.corpusthomisticum.org/.
    ${ }^{3}$ http://htl2.linguist.jussieu.fr:8080/CGL/index.jsp?link=historic.

[^8]:    ${ }^{4}$ Present perfect tense, indicative mood, first person singular of the verb moneo 'admonish'.
    ${ }^{5}$ Genitive or dative singular or nominative plural of the noun puella 'girl'.
    ${ }^{6}$ http://www.perseus.tufts.edu/hopper/morph.jsp.
    ${ }^{7} \mathrm{http}: / /$ webilc.ilc.cnr.it/ ruffolo/lemlat/index.html.
    ${ }^{8}$ http://archives.nd.edu/words.html.

[^9]:    ${ }^{9}$ http: //itreebank.marginalia.it/.

[^10]:    ${ }^{10}$ http://www.cipl.ulg.ac.be/bdlasla/.

[^11]:    ${ }^{11} \mathrm{http}: / / \mathrm{www} . l r z-m u e n c h e n . d e / ~ r a m m i n g e r / ~$
    ${ }^{12} \mathrm{http}: / /$ multiwordnet.itc.it/english/home.php.

[^12]:    ${ }^{1}$ Available at http://www.uni-stuttgart.de/lingrom/stein/forschung/ontovit/iperverb/00INDEX.html.

[^13]:    ${ }^{2}$ This expression refers to missing elements which "can be understood given conventions of interpretation, but there is no need to retrieve or construct a specific discourse referent" (Ruppenhofer et al., 2006, p. 34). For example, the annotated sentence And I only hope that in the end our roads will become so clogged with all these fume-belching cars and lorries that everyone ${ }_{T h e m e}$ will give them up and start GOING [by train] Carrier again (taken from the list of annotated sentences for this frame in FrameNet) shows a INI for Goal.
    ${ }^{3}$ This expression refers to missing elements which are "already understood in the linguistic or discourse context" (Ruppenhofer et al., 2006, p. 34). For example, the annotated sentence We Theme 'll GO [in my car] Carrier, and you can navigate (taken from the list of annotated sentences for this frame) shows a DNI for Goal.
    ${ }^{4}$ http://framenet.icsi.berkeley.edu/.

[^14]:    ${ }^{5}$ In section 4.6 .1 we will five a slightly different definition of SCFs, which is better suited for the case of Latin.

[^15]:    ${ }^{6} \mathrm{http}: / /$ centridiricerca.unicatt.it/circse_78.html.

[^16]:    ${ }^{7}$ In this column, the arrows over the arcs indicate the dependencies, pointing from the dependent to the head.

[^17]:    ${ }^{8}$ See section C. 2 for an explanation of the $\chi^{2}$ test.

[^18]:    ${ }^{9}$ In the row for the 'Atv'/'AtvV' tag, 'complement' refers to NPs like leati in Galli laeti in castra pergunt 'The Gauls happily enter the camp', which agrees with Galli. Cfr. praedicativa as defined in Pinkster (1990).

[^19]:    ${ }^{10}$ Also available at http://daphne.perseus.tufts.edu/docs/guidelines.pdf.
    ${ }^{11}$ In the text we will refer to the nodes in these trees by using the lemma of the word followed by its rank. For example, node cum-1 in figure 4.4 refers to the first node in the syntactic tree represented in figure 4.4.

[^20]:    ${ }^{12}$ In sentence (7) the infinitive esse depending on existimavit- 5 and governing dubitandum-3 is elided. Hence, the tag 'Pnom_ExD0_Obj' assigned to dubitandum indicates that this node has a 'Pnom' relation with the first (index ' 0 ') elided node of the sentence, which would have received the ' Obj ' tag if it had been present in the sentence. The same holds for quin-6 ('AuxC_ExD0_Obj').
    ${ }^{13}$ The LDT and the IT-TB annotate ablative absolutes as subordinate clauses, therefore their predicate (i.e. the participle, when it is there) depends on the main verb as an adverbial and the ablative NP (noun or pronoun, etc.) depends on the participle as subject. Cfr. the ablative absolute re frumentaria provisa in figure 4.8, where the participle provisa-3 depends on the main verb pervenit- 13 via the coordination que-6 (see section 4.5.2.2 for the annotation of coordination), and the ablative noun re-1 depends on provisa as a subject. In such cases, the fact that the subject is expressed with an ablative depends on the role of the clause it is embedded in and not on its own syntactic role with respect to its predicate. Therefore, these "ablative subjects" can be easily highlighted through their morphological labels ' $\mathrm{Sb}[\mathrm{abl}$ ' ' in the SC lexicon we built (see section A.3).
    ${ }^{14}$ Relative clauses are annotated assigning 'Atr' to the predicate of the relative clause and 'Sb' or 'Obj' to the relative pronoun, depending on its syntactic function. For example, in figure 4.6, the predicate erant-10 of the relative clause qui finitimi Belgis erant is attached to the noun it refers to (Gallis-6) as 'Atr', whereas qui-7 is the subject of the relative clause and depends on erant-10 as 'Sb'.
    ${ }^{15}$ In all accusative + infinitive constructions, the infinitive verb is tagged as 'Obj' and depends on the verb introducing the construction; the accusative NP is tagged as ' Sb ' and depends on the infinitive. For example, in figure 4.4, the verb dixeramus-34 introduces the accusative + infinitive construction quam tertiam esse Galliae partem. Therefore, the predicate of this clause (esse-31) depends on dixeramus34 via an 'Obj' relation. As in ablative absolutes, the subject of an accusative + infinitive clause is expressed with an accusative because of the role of the clause it is embedded in with respect to the main clause, rather than because of its syntactic role with respect to the predicate. This is the meaning of the "accusative subjects" ' $\mathrm{Sb}[\mathrm{acc}]$ ' occurring in the SC lexicon we built (see section A.3)

[^21]:    ${ }^{16}$ The measure for the overall lexical richness for all the words in the corpora follows the same trend: $\frac{\left|V_{T f}\right|}{|T|}=\frac{14021}{53143}=0.26, \frac{\left|V_{T l}\right|}{|T|}=\frac{5986}{53143}=0.11($ LDT $)$ and $\frac{\left|V_{T f}\right|}{|T|}=\frac{5183}{54878}=0.09, \frac{\left|V_{T l}\right|}{|T|}=\frac{1873}{54878}=0.03$
    (IT-TB). (IT-TB).
    ${ }^{17}$ In the LDT there are also 5 verbal occurrences for which the voice feature is not annotated: this is why the sum of the active and passive verbal forms does not coincide with the total number of the verbal tokens. In the second row of table 4.8 (referring to the IT-TB), we included the deponent verbs

[^22]:    in the 'active' column. The number of tokens for deponent verbs in the IT-TB is 240 . See page 69 for a discussion of the annotation of voice in the treebanks.
    ${ }^{18}$ More precisely, the type of SCC dealt with here is $S C C_{2}$. For the definition of $S C C_{2}$, see section A. 3 .
    ${ }^{19}$ See section C. 2 for an illustration of this test.

[^23]:    ${ }^{20}$ According to the annotation guidelines, the comma is considered a coordinating element; its frequency with the 'Coord' tag is the second highest among the lexical items occurring as path nodes in the LDT (482).

[^24]:    ${ }^{21}$ See section 3.2.3.1 from a description of these treebanks.
    ${ }^{22}$ For parsing experiments on Latin, see Bamman and Crane 2008; Passarotti and Ruffolo 2009.

[^25]:    ${ }^{1}$ Conversely, word-based methods aim at extracting collections of words "that co-occur significantly in the syntactic context of the studied word" (Ribas, 1994, p. 769). For this reason, word-based methods do not perform the generalization step. Here, we will focus on class-based methods, which also offer a way to resolve Word Sense Disambiguation (WSD) tasks, as we will see in section 5.4.

[^26]:    ${ }^{2}$ This implies that we will suppose that figure 5.1 approximates a portion of the WN structure; in particular, the leaf nodes in figure 5.1 (corresponding to synsets) only contain one lemma each. Notice that bird and insect appear twice in the hierarchy, once at the leaf level and once at the first parent node level.
    ${ }^{3}$ We assigned a name in capital letters to the non-leaf nodes (e.g. BIRD $=\{b i r d\}$ at the first parent node level), to distinguish them from the leaf nodes (e.g. \{bird\} at the leaf node level), since they are different as nodes, even though they may contain the same elements.

[^27]:    ${ }^{4}$ See section B.1.1 for a definition of conditional probability.
    ${ }^{5}$ For a definition of prior and posterior probability, see section B.1.

[^28]:    ${ }^{6}$ See section B.3.3 for a more precise definition of 'thesaurus tree' and 'tree cut model'.

[^29]:    ${ }^{7}$ An example of inference rule is $X$ is charged by $Y \Longrightarrow Y$ announced the arrest of $X$. This rule can be associated to the inferential SP requiring that $X$ is a Person and $Y$ is a Law Enforcement Agent or a Law Enforcement Agency; although this requirement is not sufficient, it helps filter out incorrect inferences.
    ${ }^{8}$ This is a toy example with very low constructed frequencies that are not likely to occur in a real task in computational linguistics; in what follows, the low figures statistically overestimate the actual differences between the observations and are proposed only with the purpose of illustration.

[^30]:    ${ }^{9}$ In figure 5.2 the coordinates of points are identified by the frequencies of $f l y$ ( $x$-axis) and buzz ( $y$-axis) occurring with the nouns they are labeled with. Note that the numbering on the $x$-axis starts from 2 instead of 0 , since all points have abscissas 2 or higher.
    ${ }^{10}$ The relative frequencies in table 5.3 were calcuated by dividing each cell in table 5.2 by the sum of the frequencies in the corresponding row. For example, the relative frequency of bee with fly is $\frac{3}{3+5}=\frac{3}{8}=0.375$.

[^31]:    ${ }^{11}$ As explained in section B.5.2, this clustering experiment was designed by setting the cosine as the distance measure and the single method as the linkage function. Depending on the level at which we decide to cut the tree, we get a different number of clusters. Cf. page 207 for an explanation of these terms.

[^32]:    ${ }^{12}$ Following the notation adopted and explained in chapter 4, '(in) $\mathrm{Obj}[\mathrm{acc}]$ ' represents an SC slot for the verb, i.e. a portion of the verb's SC frame containing an argument expressed with the accusative case and dependent (in the sense of Dependency Grammar) on the preposition in.

[^33]:    ${ }^{13}$ The synonimity between descendo and cado is meant here in WN terms as the fact that these two verbs share a synset. The synset for descendo actually does not appear in the original LWN and was added through the procedure explained in section 5.3.1.1.
    ${ }^{14}$ For reasons of space, in figures 5.6 and 5.7 we only represented the top nodes to which the various synsets of fons, homo, pars, solium and terra belong. The full set of synsets for these nouns, together with their corresponding synsets in English and the glosses, is listed in section B.7. The 'tub' and the 'coffin' senses of solium do not appear in LWN and hence will not accounted for here.
    ${ }^{15}$ In section 5.3 .1 .1 we will specify what we mean by 'semantic similarity' and 'closeness between syntactic patterns'.

[^34]:    16'A' stands for 'Active', 'Sb' for 'subject', 'Obj' for 'object', 'nom' for 'nominative', 'acc' for 'accusative', and 'abl' for 'ablative'. Unlike the standard adopted in chapter 4, for reasons of clarity here we incorporated the verb's lemma inside the SC structure.

[^35]:    ${ }^{17}$ Among the semantic features in table 5.5 we only reported the hypernyms of the English synset corresponding to synset number 11 of caelum in LWN, referring to the sense "outer space as viewed from the earth". For homo, we reported the hypernyms of its sixth synset "an adult male person (as opposed to a woman)". For descendo, we reported the English synset corresponding to the Latin synset \{cado, defluo, degredior, delabor, demeo\}.
    ${ }^{18}$ See chapter 1 for a more detailed discussion of these points, within an overview on Latin syntax and the argument structure of Latin verbs.

[^36]:    ${ }^{19}$ This figure corresponds to the size of LWN in May 2009.
    ${ }^{20}$ This operation could theoretically be performed automatically by using an onomastic lexicon or Named-Entity Recognizer, a tool that exists for a number of modern languages (cf. Dell'Orletta et al. 2007 for Italian, for example), but not for Latin yet.

[^37]:    ${ }^{21}$ This phenomenon is called "omografia radicale" ('root omography') in Passarotti and Ruffolo (2004) and includes the cases where two etymologycally different lemmas happen to display the same form. For example, in frons this identity manifests itself at the nominative and vocative singular forms.
    ${ }^{22}$ For impedio/impello, in table 5.6 we listed the forms found in the two treebanks (inpello in LDT and impedio in IT-TB), while LWN contains both impedio and impello. Similarly, in the IT-TB we find affigo but no form of affligo, as we do in the LDT.
    ${ }^{23}$ The synset for descendo in table 5.7 does not appear in the original LWN. We added it through the procedure explained in the next paragraph.

[^38]:    ${ }^{24}$ The Italian lemma uscio is not present in MWN.

[^39]:    ${ }^{25}$ In case there are more than one such construction, the first construction realising the maximum is chosen.
    ${ }^{26}$ Bayes' theorem is sketched in Appendix B.1.2.

[^40]:    ${ }^{27}$ The full sets of semantic properties for terra and homo are displayed in sections B.7.2.1 and B.7.2.2.

[^41]:    ${ }^{28}$ In order to partially remedy to the problem of zero frequencies, we adopted a smoothing in the probability calculations, as explained in Appendix B.8.1.

[^42]:    ${ }^{29} \mathrm{We}$ only reported the constructions containing the slot '(in)A_Obj[acc]'.
    ${ }^{30}$ We remind that 'NP' refers here to proper nouns of persons.

[^43]:    ${ }^{31}$ This statement is confirmed by measures of statistical dispersion, such as the Interquartile Range and the Median Absolute Deviation. For the values of these measures on our data, see section B.9.0.1.

[^44]:    ${ }^{32} \mathrm{http}: / /$ www.perseus.tufts.edu/hopper /.
    ${ }^{33}$ www.corpusthomisticum.org/.
    ${ }^{34}$ www.thelatinlibrary.com/.

[^45]:    ${ }^{35}$ These data were given to us by David Bamman and Paolo Ruffolo, respectively. See Bamman and Crane (2008); McGillivray et al. (2009) for descriptions of the parsing exsperiments and section 3.2.3.2 for an introduction to parsed Latin corpora.

[^46]:    ${ }^{1}$ By the term 'adposition' I refer to the common class to which prepositions and postpositions belong.
    ${ }^{2}$ From now on, we will call these elements 'adverbial elements'.
    ${ }^{3}$ Luraghi (2003) argues against the equivalence between order of elements and constituency: "contiguity of sub-constituents does not constitute a diagnostic for individuating the head of a phrase".
    ${ }^{4}$ For a discussion on the role of these particles in ancient Greek in relation with adverbs and prepositions, see Viti (2008), Haug (2009) and Luraghi (2010b).

[^47]:    ${ }^{5}$ This differentiation could be defined as 'reanalysis': cf. the definition in Langacker (1977, p. 58) "change in the structure of an expression or class of expressions that does not involve any immediate or instrinsic modification of its surface manifestation".

[^48]:    ${ }^{6}$ As pointed out by Traugott (2006), here "construction is used in a pre-theoretical way", although Construction Grammar (Goldberg, 1995; Fillmore et al., 2003) and other theories are also relevant in this context.

[^49]:    ${ }^{7}$ For the synchronic sense of lexicalization used in Talmy's theory, see section 6.3.3.
    ${ }^{8}$ For a discussion on the categories of lexical words and grammatical words, see e.g. Brinton and Traugott (2005, pp. 11-18).

[^50]:    ${ }^{9}$ However, it has been shown that some preverbs emerged in Late Latin restricted to the technical domain. Urso (1999) reports a relatively high number of newly-formed PVs attested in two medical texts - Gynaecia and Passiones - by Caelius Aurelianus (5th century A. D.), who traslated them from Greek originals by Soranus of Ephesus (2nd century A. D.). Moreover, in their paper on preverbs and particles in Old French, Dufresne et al. (2003) show that "aspectual prefixes were very productive in Old French" but no longer in Modern French, where some PVs lexicalized; they note that the productivity of preverbs stopped around the end of the 16 th century.
    ${ }^{10}$ For an alternative interpretation of the spatial meaning of these prepositions, see Garcìa Jurado (1991).
    ${ }^{11}$ Actually, the prefix $a$ - in Italian can be used to express the shift to a specific state as in the parasynthetic verb abbellire 'become more attractive', formed by bello 'attractive', the prefix $a$ - and the suffix -ire (Rohlfs, 1970, p. 347).
    ${ }^{12}$ A similar change happened from Old English to Middle English, where the preverbs were gradually replaced by postverbal particles used in phrasal verbs (Brinton, 1988; Imbert, 2008).

[^51]:    ${ }^{13}$ For a discussion on lexicalization, see section 6.3.2.
    ${ }^{14}$ 'Verbi sintagmatici' are defined in Simone (1997, p. 49) as "sintagmi formati da una testa verbale e da un complemento costituito da una "particella" (originariamente un avverbio) uniti da una coesione sintattica di grado elevato al punto che non si può commutare il VS intero con una sola delle sue parti" 'phrases composed of a verbal head and a complement contituted by a "particle" (originally an adverb) joined by such a strong syntactic cohesion that the verb cannot be replaced by one of its parts'.

[^52]:    ${ }^{15}$ However, it is not clear which simple verb the preverb intro- was prefixed to in intrare.

[^53]:    ${ }^{16}$ Lehmann (1983, p. 156) speficies that "if the relatum of the preverb is the object of the verbum compositum, then the locatum is either in the ablative or not adjoinable".
    ${ }^{17}$ Bortolussi (2005) analyses these cases, differentiating the more frequent presence of the accusative case for bi-valent verbs (such as adeo 'go to') and the rarer presence of the accusative case for trivalent verbs (such as transeo 'go across'). He also distinguishes this specific construction from the more generale lative accusative, used with any motion verb and typically with proper nouns to express direction towards a place; this brings him to note that the double accusative constructions are only present when a physical movement is implied: cf. the sense 'lead across' of traduco vs. the sense 'translate'.

[^54]:    ${ }^{18}$ Traina (1955) analyses special types of appositional NPs such as Antiochiae . . . celebri quondam urbe et copiosa 'in Antioch, once populous and wealthy city', where an adjective is added to the proper NP. According to Traina, in these cases the preposition is absent when the apposition gives additional information for determining the proper NP, while the preposition is present when the apposition "completes the sense of the proposition". This remark underlines the role of non-diachronic factors affecting the presence of construction 1 (CasePrev), which enriches the picture given by the diachronic hypotheses presented in this section.

[^55]:    ${ }^{1}$ We chose the edition used in the Perseus Digital Library; this provided us with a freely available digital text.

[^56]:    ${ }^{2}$ In the following list we summarised, for each preverb, the senses given by the Lewis and Short (1879) dictionary.

[^57]:    ${ }^{3}$ This criterion regards picking only those verbs prefixed with a preverb belonging to a specific list, among all verbs in the corpora. Since this is a morphological criterion, it is in theory possible to perform it in a semi-automatic way by first using a lemmatizer, such as LEMLAT (Passarotti and Ruffolo, 2004): cf. section 3.2.2.1. The lemmatizer would produce "candidate" PVs prefixed with a specific preverb; then we would have to manually check for PVs that are not captured by this procedure.
    ${ }^{4}$ For a survey on annotated corpora and computational resources for Latin, see chapter 3.
    ${ }^{5}$ For an overview on parsers for Latin, see section 3.2.3.2.
    ${ }^{6}$ In this class we included also verbs implying a movement while not being motion verbs, as noted by Bortolussi (2005, p. 285) about circum- 'around'. Consequently, in addition to verbs like adduco 'lead to', we considered verbs like circumdo 'put around'.

[^58]:    ${ }^{7}$ For the three studies we carried out, we will refer to several subsets taken from these data. Table 7.1 only reports the size of the original data from which these subsets were extracted.

[^59]:    ${ }^{8}$ For a similar analysis on the alternation between construction 1 (CasePrev) and 3 (PrepPrev) in a smaller dataset see Meini and McGillivray (2010).

[^60]:    ${ }^{9} \mathrm{~A}$ MCA on the same dataset replacing 'era' with 'author' yields a lower percentage of inertia explained by the first two axes (42.4\%).

[^61]:    ${ }^{1}$ For each item in this list, the corresponding attribute for the first word occurring in table A. 1 (i. e. quem) is given in brackets.
    ${ }^{2}$ The number 1 following the lemma quis refers to the first sense of this lemma in the Lewis and Short (1879) dictionary.

[^62]:    ${ }^{3}$ More details on the morphological tagset used in the LDT are available in the "readme.txt" file at http://nlp.perseus.tufts.edu/syntax/treebank/.
    ${ }^{4}$ For each item in this list, we give in brackets the corresponding instance for the first word in table A.2, praedicatum.

[^63]:    ${ }^{5}$ By 'nominals' we mean those PoSes that are inflected with case and gender, such as nouns, adjectives and pronominals.
    ${ }^{6}$ By 'participial' we mean verbal forms provided with nominal inflection, such as participles, gerundives and gerunds.
    ${ }^{7}$ 'Formal variation' here refers to those graphical variations inserted bu Thomas for stylistic purposes.
    ${ }^{8 \prime}$ Graphical variation' refers to those graphical variations due to Thomas' usus scribendi, to scribes or editors, but not to stylistic reasons. For a description of the tagset of the IT-TB, see http://gircse.marginalia.it/\%7Epassarotti/tagset/IT\%20tagset.pdf .

[^64]:    ${ }^{9}$ For reasons of clarity, in representing table A. 3 we replaced the word IDs with their lexical forms.

[^65]:    ${ }^{10}$ See page 66 ff . for the details of the annotationd of these tags.

[^66]:    ${ }^{11}$ Following the tokenization convention of the LDT, the que enclitic is tokenised as a separate word and follows the word it is linked to.

[^67]:    ${ }^{12}$ For reasons of clarity, in table A. 5 we replaced the word IDs with the wordforms in the sentence, and we recorded the functional label of each word.
    ${ }^{13}$ For reasons of clarity, in table A. 6 we replaced the word IDs with the wordforms in the sentence.

[^68]:    ${ }^{15}$ The same correlation is observed when the frequencies are plotted on a log scale which corrects for outliers.
    ${ }^{16} p$-values are adjusted for multiple testing with the Bonferroni correction.

[^69]:    ${ }^{1}$ A probability mass function $f_{X}$ of a discrete random variable $X$ is a function that assigns the probability that $X$ has a certain value: $f_{X}(x)=P(X=x)$.

[^70]:    ${ }^{2}$ We assigned a name in capital letters to the non-leaf nodes such as ANIMAL $=\{$ animal $\}, \operatorname{BIRD}=\{$ bird $\}$ and INSECT $=\{$ insect $\}$.
    ${ }^{3}$ All logarithms are base two unless otherwise stated.

[^71]:    ${ }^{4}$ In this particular example, the frequency of each class coincides with the frequency of the noun it contains because of the particular composition of the classes (each of them contains only one noun, which does not appear in any other class), the only exception being represented by bird, which occurs in two classes.

[^72]:    ${ }^{5}$ This example and the following data are taken from Li and Abe (1998).

[^73]:    ${ }^{6}$ This implies that a word node is true if the word is a possible argument for the verb; otherwise, it is false.

[^74]:    ${ }^{7}$ This distinguishes clustering from classification, where the groups forming the input are known in theory, before the analysis can support them or not. Cf. Baayen (2008, pp. 148 ff .) for an illustration of classification.
    ${ }^{8}$ In geometry a distance or metric on a set $S$ is a function that assigns a non-negative real number $d(x, y)$ to each pair $(x, y) \in S$ such that the following three conditions are met for each $x, y, z \in S$ : $d(x, y)=0$ if and only if $x=y ; d(x, y)=d(y, x) ; d(x, z) \leq d(x, y)+d(y, z)$.
    ${ }^{9}$ The $n$-dimensional version of (B.2) is $d(u, w)=\sqrt{\sum_{i=1}^{n}\left(u_{i}-w_{i}\right)^{2}}$ for two vectors $u$ and $w$ whose coordinates are $\left(u_{1}, \ldots, u_{n}\right)$ and $\left(w_{1} \ldots, w_{n}\right)$, respectively.
    ${ }^{10}$ The $n$-dimensional version of (B.3) is $\cos (u, w)=\frac{\sum_{i=1}^{n} u_{i} w_{i}}{\sqrt{\sum_{i=1}^{n} u_{i}^{2} \sum_{i=1}^{n} w_{i}}}$ for two vectors $u=$ $\left(u_{1}, \ldots, u_{n}\right)$ and $w=\left(w_{1}, \ldots, w_{n}\right)$.

[^75]:    ${ }^{11}$ This points at one of the aknowledged disadvantages of hierarchical clustering, i.e. the fact that even unrelated objects appear eventually joined by the algorithm.

[^76]:    ${ }^{12}$ Note that, by convention, cluster 8 indicates the new cluster created by the grouping of crow and eagle, while the single clusters previously containing these nouns are now empty.
    ${ }^{13}$ Both the cluster analysis and the dendrogram were produced using the functions pdist, linkage

[^77]:    and dendrogram in MATLAB ${ }^{\circledR}$ (MathWorks Inc., 2004).
    ${ }^{14}$ The inconsistency coefficient measures the relative consistency of each link in a hierarchical cluster tree and is computed by comparing the height of a link in a cluster hierarchy with the average height of links below it. Low inconsistency coefficients correspond to links joining distinct clusters, while high inconsistency coefficients correspond to links joining indistinct clusters. Inconsistency coefficients, as well as other measures for evaluating a clustering can be calculated in MATLAB through specific functions; this will not be treated here.
    ${ }^{15}$ This example is taken from McCarthy (2001, p. 17).

[^78]:    ${ }^{16}$ This example is taken from Erk (2007).
    ${ }^{17}$ The inverse document frequency is defined as $w t_{r}(w)=\log \left(\frac{\text { num.words }}{\text { num.wordstowhose context } w \text { belongs }}\right.$.

[^79]:    ${ }^{18} \mathrm{http}: / /$ multiwordnet.fbk.eu/online/multiwordnet.php.

[^80]:    ${ }^{19}$ For a definition of median see the next paragraph.

[^81]:    ${ }^{20}$ The results of a binomial test on 4 successes gives the following $95 \%$ confidence interval and estimated probability of success: $[0.028,0.237]$ and 0.1 .

[^82]:    ${ }^{21}$ The number of significant cases at the 0.01 level was 12 .

[^83]:    ${ }^{1} \mathrm{~A} \chi^{2}$ distribution is defined as "the square of the deviation of a score from its population mean divided by the population deviance" (Hinton, 2004, p. 248). We will not go into the details of this definition, for which we refer to Gorard (2003) and Hinton (2004), for example.

[^84]:    ${ }^{2}$ The input of a $\chi^{2}$ test is identified by the number of degrees of freedom, calculated as $(n r-1)(n c-$ $1)=1$ in our example. The degrees of freedom count how many values in the table we need to know to infer the values of the whole table, given the row and column totals. Once we know the number of degrees of freedom and a significance level, we can calculate the corresponding critical value for the $\chi^{2}$ statistic: if this value is greater than the value we calculated, then we can reject the null hypothesis at the decided level of significance, e. g. 0.05. Tables of critical values are usually recorded in statistics manuals. Here, we use the table in Hinton (2004, p. 370).

[^85]:    ${ }^{3}$ The normal distribution or Gaussian function is a probability distribution represented by the wellknown bell shape and characterized by the fact that most of its values tend to cluster around the mean (the peak of the bell). Its wide use is also explained by the fact that the frequency distribution of the means of a large number of equally-sized samples tends to become a normal distribution (central limit theorem). See e.g. Hinton (2004, pp. 28-33).

[^86]:    ${ }^{4}$ A binomial distribution describes the distribution of a number of independent experiments yielding only two possible outcomes, called Bernoulli outcomes.

[^87]:    ${ }^{5}$ This is the standard R notation for binary variables, for which the values are defined as ' 0 ' and ' 1 '. For sake of clarity, in this dissertation we will assign more meaningful names to the values. For 'construction', for example, we named the values as 'CasePrev' and 'Preposition'.

[^88]:    ${ }^{6}$ The rank of a matrix is the maximal number of linearly independent rows (or equivalently columns) of $A$.
    ${ }^{7}$ The transpose of $V$ is the matrix whose rows are the columns of $V$ : that is, its $(i, j)$-th element is $V(j, i)$.

[^89]:    ${ }^{8}$ The identity matrix is the matrix that has 1 along its diagonal and 0 everywhere else.
    ${ }^{9}$ In fact, this case is particularly simple, since the rank of the matrix is 2, given that the number of columns is 3 . In general, the number of principal axes cannot be larger than the rank of the matrix $k \leq \min (r-1, c-1)$.

[^90]:    ${ }^{10}$ 'Era' has the value 'Early' for Plautus, 'Classical' for Caesar, Cicero, Ovid, Petronius, Propertius, Sallust, and Vergil, and 'Late' for Jerome and Thomas.
    ${ }^{11}$ A CFA on the variables 'animacy', 'preposition' and 'author' is not meaningful because many observed frequencies fall under the recommended threshold of 5 .

