

Parsimonious Vole

A Systemic Functional Parser for English



Universität Bremen

Eugeniu Costetchi

Supervisor: Prof. John Bateman

Advisor: Dr. Eric Ras

Faculty 10: Linguistics and Literary Studies
University of Bremen

This dissertation is submitted for the degree of
Doctor of Philosophy

May 2019

thank you

for Adriana

Tamara, Ion and Cristi Costetchi

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

Eugeniu Costetchi

May 2019

Acknowledgements

This thesis owes much to the many people who have guided me, supported me, and inspired me throughout the preparation and writing of this work. Below, I attempt to list some of these colleagues, family, and friends, but I cannot hope to thank everyone by name. Thus, upfront, to each and every one I offer my heartfelt thanks.

First and foremost I shall forever be grateful to my academic supervisor, John Bateman, for giving me his guidance, deep insight, patience and a lot of diligent proofreading work when the time finally came. Without him I would never have become a computational linguist and this thesis would not have happened. Similarly, I am very much indebted to my advisor, Eric Ras, for his kind encouragements, guidance and support right from the beginning starting with PhD proposal writing. Without Eric this thesis could not have started in the first place.

I believe knowledge is created between people and I would like to thank all those who have shared this process with me. To everyone who has shared a chat over coffee, a talk around the table or a talk at a conference or seminar, thank you. In particular I would like to thank my colleague and friend Muriel Foulonneau, who invited me to join LIST research centre and was always engaging in stimulating discussions. Thanks to Anke Schulz who was the first person I met in real need of an SFL parser because she was performing, at that time, tedious manual corpus annotation. That corpus annotation later became part of the Parsimonious Vole evaluation. Thank you to Ela Oren for the work we have done together on corpus annotation, also used in the parser evaluation; for inviting me on a short scientific mission to Tel Aviv University; and from whom I have learned about Obsessive Compulsive Disorder. My deep gratitude goes to Daniel Couto Vale who always welcomed me in Bremen and enthusiastically shared his knowledge on Systemic Functional Linguistics.

There are many friends that have shared this experience with me and I can't thank each of them enough. To any I have inadvertently left out please don't think you are forgotten. A big thank you goes to my friend Mikolaj Podlaszewski with whom I shared lots of thought-provoking philosophical discussions, sometimes fierce debates and who provided me with lots of constructive criticisms. My friend, Andrei Mihalceanu, who

unfortunately passed away, has my sincere gratitude for enthusiastic philosophical discussions on language, mind, determinism and entropy. Thanks to Christoph Stahl for his friendship, encouragement and for putting me always to work on my thesis.

Even more so than friends, family are there in person and spirit when you need them most, and that is why they deserve the greatest gratitude of all. A huge thank you to my parents Tamara and Ion Costetchi for unconditional love, encouragement and support. I want to thank my younger brother Cristi. I haven't always been the best big brother for him, but he has always been there for me. But most of all, I thank Adriana, my beloved wife, who sometimes pushed hard and some other times gently encouraged me in the last phase of this thesis, and then patiently waited for the manuscript to mature. This thesis is for her.

Finally, I gratefully acknowledge the support of Luxembourg National Research Fund through an AFR PhD grant which made this work possible in the first place. I also want to thank all those who gave me feedback on drafts along the way. However, mistakes, be them of the conceptual or typographic variety, remain mine and mine alone.

Abstract

Building a natural language parser can be seen as a task of creating an artificial text reader that understands the meaning expressed in some text. This thesis aims at a reliable modular method for parsing unrestricted English text into feature-rich constituency structure using Systemic Functional Grammars (SFG). SFGs are chosen because of their versatility to account for the complexity and phenomenological diversity of human language.

The descriptive power of a Systemic Functional Grammar (SFG) lies in its separation of descriptive work across “structure” (i.e., syntagmatic organisations) and “choice” (i.e., paradigmatic organisations). A shortcoming, however, is that SFL has been primarily concerned with the paradigmatic axis of language, and accounts of the syntagmatic axis of language, such as the syntactic structure, have been put in the background.

Moreover, parsing with features that depart from directly observable grammatical variations towards increasingly abstract semantic features comes at the cost of high computational complexity, which still presents today the biggest challenge in parsing broad coverage texts with full SFGs. Previous research has discussed how each successive attempt to construct parsing components using SFL then led to the acceptance of limitations either in grammar size or in language coverage.

One of the main contributions of this thesis is the investigation to what degree cross-theoretical bridges can be established between Systemic Functional Linguistic (SFL) and other theories of grammar, Dependency Grammar in particular, in order to compensate for the limited syntagmatic accounts. A second main contribution is to research how suitable predefined graph patterns are for detecting systemic features in the constituency structure in order to reduce the complexity of identifying increasingly abstract grammatical features.

The technical achievement of this thesis lies in the development and evaluation of a SFG parser, named Parsimonious Vole. The implementation follows a pipeline architecture comprising of two major phases: (a) creation of the constituency structure

from Dependency graphs and (b) structure enrichment with the systemic features using graph pattern matching techniques.

The empirical evaluation is based on two manually annotated corpora. First, covers constituency structure and Mood features and, second, covers the more abstract Transitivity features. The parser accuracy at generating constituency structure (76%) is slightly lower than that achieved in previous research, while the accuracy to detect Mood (60%) and Transitivity (42%) could not be compared to any previous works because either such features are missing or results are not comparable.

The current work concludes that (a) reusing parse results with other grammars for structure creation and (b) employing graph patterns for enrichment with systemic features constitutes a viable solution to create feature-rich constituency structures in SFL style.

Table of contents

1	Introduction	1
1.1	On artificial intelligence and computational linguistics	1
1.2	Living in a technologically ubiquitous world	3
1.3	NLP for business	4
1.4	Linguistic framework	5
1.5	A systemic functional analysis example	8
1.6	Challenges of parsing with SFGs	13
1.6.1	Syntagmatic descriptions in SFL	14
1.6.2	Computational complexity appears in parsing	15
1.6.3	Parsing with semantic features	17
1.6.4	Covert elements	18
1.6.5	Problem summary	20
1.7	Goals and scope of the thesis	21
1.7.1	On theoretical compatibility and reuse	22
1.7.2	Towards the syntagmatic account	23
1.7.3	Towards the paradigmatic account	24
1.7.4	Parsimonious Vole architecture	26
1.8	Thesis overview	29
2	An overview of selected work on parsing with SFG	33
2.1	Winograd’s SHRDLU	34
2.2	Kasper	34
2.3	O’Donnell	35
2.4	O’Donoghue	36
2.5	Honnibal	37
2.6	Summary	38

3	Systemic functional theory of grammar	39
3.1	A word on wording	40
3.2	Sydney theory of grammar	43
3.2.1	Unit	44
3.2.2	Structure	46
3.2.3	Class	47
3.2.4	System	48
3.2.5	Functions and metafunction	51
3.2.6	Lexis and lexico-grammar	53
3.3	Cardiff theory of grammar	53
3.3.1	Class of units	54
3.3.2	Element of structure	55
3.3.3	Item	56
3.3.4	Componence and obscured dependency	57
3.3.5	Filling and the role of probabilities	58
3.4	Critical discussion of both theories: consequences and decisions for parsing	60
3.4.1	Relaxing the rank scale	60
3.4.2	Approach to structure formation	63
3.4.3	Relation typology in the system networks	63
3.4.4	Unit classes	64
3.4.5	Syntactic and semantic heads	67
3.4.6	Coordination as unit complexing	69
3.5	Summary	75
4	Parsimonious Vole grammar	77
4.1	Grammatical units	77
4.1.1	Verbal group and clause boundaries	77
4.1.2	Clause	79
4.1.3	Nominal Group	80
4.1.4	Adjectival and Adverbial Groups	84
4.2	System networks	87
4.2.1	MOOD	87
4.2.2	TRANSITIVITY	90
4.2.3	Process Type Database	92
4.3	Summary	93

5	Dependency grammar (DG)	95
5.1	Origins of dependency theory	95
5.2	Evolution into modern dependency theory	101
5.2.1	Definition of dependency	101
5.2.2	Grammatical function	102
5.2.3	Projectivity	103
5.2.4	Function words	103
5.3	Dependency grammar in automated text processing	105
5.4	Stanford dependency model	107
5.5	Stanford dependency representation	109
5.6	Cross theoretical bridge from DG to SFG	110
6	Government and Binding Theory (GBT)	117
6.1	Introduction to GBT	118
6.1.1	Phrase structure	119
6.1.2	Theta theory	121
6.1.3	Government and Binding	123
6.2	On Null Elements	126
6.2.1	PRO Subjects and control theory	127
6.2.2	NP-traces	129
6.2.3	WH-traces	131
6.3	Placing Null Elements into the Stanford dependency grammar	133
6.3.1	PRO subject	133
6.3.2	NP-traces	137
6.3.3	Wh-traces	139
6.3.4	Wh-traces in relative clauses	142
6.4	Discussion	144
7	Graphs, Feature Structures and Systemic Networks	145
7.1	General definitions	146
7.2	Graph traversal	152
7.3	Pattern graphs	154
7.4	Graph matching	158
7.5	Pattern based operations	163
7.5.1	Pattern based node update	164
7.5.2	Pattern based node insertion	166
7.6	Systems and Systemic Networks	167

7.7	On realisation rules	171
7.8	Summary	174
8	Creating the systemic functional constituency structure	175
8.1	Canonicalisation of dependency graphs	175
8.1.1	Loosening conjunction edges	176
8.1.2	Transforming copulas into verb centred clauses	177
8.1.3	Non-finite clausal complements with adjectival predicates (a pseudo-copula pattern)	179
8.2	Correction of errors in dependency graphs	181
8.2.1	Free prepositions and <i>prep</i> relation	181
8.2.2	Non-finite clausal complements with internal subjects	182
8.2.3	First auxiliary of non-finite POS	182
8.2.4	Prepositional phrases as false prepositional clauses	183
8.2.5	Mislabelled infinitives	183
8.2.6	Attributive verbs mislabelled as adjectives	184
8.2.7	Non-finite verbal modifiers with clausal complements	184
8.2.8	Demonstratives with a qualifier	185
8.2.9	Topicalised complements labelled as second subjects	187
8.2.10	Misinterpreted clausal complement of the auxiliary verb in inter- rogative clauses	188
8.3	Creation of systemic constituency graphs from dependency graphs . . .	188
8.3.1	Dependency nature and implication on head creation	189
8.3.2	Tight coupling of dependency and constituency graphs	190
8.3.3	Rule tables	191
8.3.4	Creating partial constituency graph through top-down traversal	194
8.3.5	Completing the constituency graph through bottom-up traversal	197
8.4	Summary	200
9	Enrichment of the constituency graph with systemic features	201
9.1	Creation of MOOD graph patterns	202
9.2	Enrichment with MOOD features	205
9.3	Creation of empty elements	208
9.3.1	PRO and NP-trace Subjects	208
9.3.2	Wh-trances	212
9.4	Cleaning up the PTDB	213
9.5	Generation of the TRANSITIVITY graph patterns	216

9.6	Enrichment with TRANSITIVITY features	220
9.7	Summary	222
10	Empirical evaluation	223
10.1	Evaluation corpus	224
10.1.1	OE corpus	225
10.1.2	OCD corpus	226
10.1.3	Differences between corpus annotation and parser output	227
10.2	Evaluation methodology	229
10.2.1	Corpus annotations as a set of mono-labelled segments	229
10.2.2	Parser output as a set of mono-labelled segments	230
10.2.3	Segment alignment method and evaluation data	232
10.3	Evaluation of syntactic structure generation	235
10.3.1	Segmentation evaluation	236
10.3.2	Unit class evaluation	240
10.3.3	Clause Mood elements evaluation	243
10.3.4	Clause Transitivity elements evaluation	244
10.4	Evaluation of systemic feature assignment	246
10.4.1	Evaluation of MOOD systemic feature assignment	246
10.4.2	Evaluation of TRANSITIVITY systemic feature assignment	250
10.5	Summary	255
11	Conclusions	259
11.1	Research questions and main findings	261
11.2	Limitations and future work	263
11.3	Practical applications	273
11.4	Final word	274
	References	275
A	SFL Syntactic Overview	291
A.1	Cardiff Syntax	291
A.1.1	Clause	291
A.1.2	Nominal Group	291
A.1.3	Prepositional Group	292
A.1.4	Quality Group	292
A.1.5	Quantity Group	292
A.1.6	Genitive Cluster	292

A.2	Sydney Syntax	293
A.2.1	Logical	293
A.2.2	Textual	293
A.2.3	Interactional	293
A.2.4	Experiential	293
A.2.5	Taxis	294
B	Stanford Dependency schema	295
C	Penn treebank tag-set	299
D	Rules for clause complex taxis analysis	301
E	Mapping dependency to constituency graph	309
F	Normalization of PTDB and Cardiff TRANSITIVITY system	313
G	A selection of graph patterns	315
H	Auxiliary algorithms	319
I	Annotation guidelines for OCD corpus	321
I.1	Constituency	321
I.2	Clause partition	323
I.3	The tricky case of prepositional phrases	324
I.4	Making selection from the MOOD system network	325
J	Empirical evaluation data	327