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 many 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 supervisor, 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 need of an SFL parser because she was performing tedious manual corpus annotation. That corpus annotation later became part of Parsimonious Vole evaluation. Thank you to Ela Oren for the work we have done together on corpus annotation, needed in 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 up with me always working 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 gently pushed and encouraged me in the last phase of this thesis and 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 which 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), which 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 to a considerable extent in its separation of descriptive work across "structure" (i.e., syntagmatic organisations) and "choice" (i.e., paradigmatic organisations). Since it was established, SFL has been primarily concerned with the paradigmatic axis of language. 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. O'Donnell & Bateman (2005) discuss how each successive attempt to construct parsing components using SFL then necessarily led to the acceptance of limitations either in grammar size or in language coverage in order to proceed.

One of the main contributions of this thesis is the investigation to what degree can cross-theoretical bridges 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. Second main contribution is investigation on how suitable are predefined graph patterns for detecting systemic features in the constituency structure, in order to reduce the complexity of identifying increasingly abstract grammatical features.

The practical 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: *creation* of the constituency structure

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

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

The current work shows 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