Univerzita Karlova
Filozofická fakulta

Disertační práce

2018 Slavomír Čéplö

Univerzita Karlova
Filozofická fakulta

Ústav srovnávací jazykovědy
Filologie – matematická lingvistika

Slavomír Čéplö

Constituent order in Maltese: A quantitative analysis

Vetný slovosled v maltčine: Kvantitatívna analýza

Disertační práce

Vedoucí práce: doc. Petr Zemánek, CSc.

2018

Acknowledgments

Researching and writing a PhD thesis is a solitary experience and yet when reflecting upon the process, I find that there are many who in one way or another contributed to it. I am indebted to all of them, namely:

- first and foremost, to my advisor Petr Zemánek who is the only reason that this dissertation cum bona, tum finita est;
- to Lonneke van der Plas, Michael Cooperson, Lameen Souag, Adam Ussishkin, Andy Wedel and Chris Lucas, all friends senior to me (not all in age, but all in wisdom and experience) who at one point or another provided guidance and support;
- to fellow travelers and friends Shiloh Drake, Matthew Scarborough and Adam Benkato for the very same;
- to Steve "languagehat" Dodson, Marie-Lucie "m-l" Tarpent, John Cowan, John Emerson, David Marjanović, Christopher Culver, AJP Korona úr and all the Hatters for the innumerable examples of erudition that inspire me and keep me going;
- to fellow mustašriqūn Adela, Katka, Emík and Jarík for listening to my ramblings and putting up with me;
- to those of my former co-workers and clients at HP whose many shortcomings inspired and forced me to overcome mine (and learn Perl);
- to those of my former co-workers at HP who have no shortcomings and who have become more, LuciaM, SoňaP, AttilaP, MarekO, BranislavM and ZuzanaS, for their support and patience;
- to Jiří Milička and Amir Zeldes for their assistance with various technical aspects of the finished work;
- to fellow members of the Ancient and Honorable Secret Society of A3net for all the flamewars where I learned a lot about programming and everything about how to structure an argument, eviscerate the opponent's and (never) admit being wrong;
- to the proprietor of and the staff at the Habibi Café and Shisha Bar where I spent many hours working on the thesis and, inter alia, came up with the ideas behind section 7.4;
- to Albert Gatt for his immeasurable help with just about every aspect of this thesis and anything I have ever done involving Maltese;
- to Tony Burke without whose support, inspiration and example I would have given up a long time ago;
- and finally, to .j, for developing PosTagger, for providing all kinds of help, encouragement and constructive criticism, and you know, in general.

None of this is anybody's fault but mine.

Vyhlásenie

Vyhlasujem, že som túto dizertačnú prácu napísal samostatne a za použitia uvedených a riadne citovaných prameňov a literatúry a že táto práca nebola použitá v rámci iného vysokoškolského štúdia alebo na získanie iného či rovnakého titulu.

V Bratislave, 15. marca 2018

Abstract

This dissertation is a quantitative analysis of constituent order (i.e. the order of subject, verb and object) in contemporary (post-2000) Maltese, a Semitic language descended from North African varieties of Arabic, spoken primarily in the Malta archipelago and Australia. The analysis is based on data collected in two corpora: a general corpus and a syntactically annotated corpus (dependency treebank); the compilation and description of the treebank is the secondary aim of this work.

The dissertation comprises 8 chapters divided into two parts: the first three chapters of which provide a conceptual foundation (chapter 1), a review of existing major approaches to the study of constituent order (chapter 2) and a review of previous works on Maltese constituent order (chapter 3). Using these as the background, chapter 4 then sets the research questions and methodology. The remaining three chapters form the core of the dissertation: chapter 5 describes the composition and enrichment of the general corpus of Maltese. Chapter 6 then provides a thorough description of the Maltese treebank and the annotation decisions, thus in effect assembling a sketch of Maltese syntax. Chapter 7 then provides the actual quantitative analysis of constituent order in Maltese based on the treebank, arriving at the conclusion that the dominant order is SVO or SV/VO and making a detailed case for rejecting previous descriptions of Maltese constituent order as "discourse-configurational", "topic-prominent" and "free". The final chapter summarizes the findings and lays out a number of avenues for further research into the topic.

Key words: Maltese, computational linguistics, syntax, dependency syntax, treebank, constituent order

Rezumé

Predkladaná dizertačná práca si v prvom rade kladie za cieľ kvantitatívne popísať vetný slovosled v maltčine. Maltčina je semitský jazyk, ktorý sa vyvinul zo severo-afrických dialektov arabčiny; hovorí sa ňou na Malte, Goze a Comine, ako aj v Austrálii, a má štatút oficiálneho jazyka Európskej únie. Kvantitatívna analýza slovosledu maltčiny v tejto práci je založená na dátach zhromaždených v dvoch korpusoch maltčiny, a to vo všeobecnom korpuse a v korpuse syntakticky anotovanom na základe princípov závislostnej syntaxe; príprava a anotácia tohto závislostného korpusu je druhým cieľom tejto práce.

Táto dizertačná práca sa skladá z ôsmich kapitol rozdelených do dvoch častí: prvé tri kapitoly popisujú konceptuálne východiská práce (kapitola 1), zhŕňajú prevažujúce koncepcie popisu vetného slovosledu (kapitola 2) a analyzujú predchádzajúce práce, ktoré sa venovali popisu vetného slovosledu v maltčine (kapitola 3). Kapitola 4 na základe predchádzajúcich kapitol stanovuje výskumné otázky a popisuje spôsob získania odpovedí na ne. Ďalšie tri kapitoly tvoria jadro práce, kde najprv kapitoly 5 a 6 opisujú zber a spracovanie údajov: kapitola 5 obsahuje popis všeobecného korpusu maltčiny, kapitola 6 zas podrobne opisuje syntakticky anotovaný korpus a proces jeho vzniku; tento proces v zásade predstavuje kompiláciu stručného popisu syntaxe maltčiny. Kapitola 7 sa venuje vlastnej kvantitatívnej analýze slovosledu v maltčine, kde záverom je, že dominantný slovosled v maltčine je podmet-prísudok-predmet, resp. podmet-prísudok + prísudok-predmet; kapitola 7 zároveň predkladá sadu argumentov, z ktorých vyplýva nevhodnosť popisu slovosledu v maltčine ako voľného, pragmaticky orientovaného alebo diskurzne konfiguračného. Posledná kapitola zhŕňa výsledky práce a zároveň naznačuje smer, ktorým by sa budúci výskum problematiky mal uberať.

Kľúčové slová: maltčina, počítačové spracovanie jazyka, syntax, závislostná syntax, syntakticky anotovaný korpus, vetný slovosled

Contents

Acknowledgments —— V

Vyhlásenie —— VII

Abstract —— IX

Rezumé —— XI

List of Tables —— XIX

List of Figures —— XXI

1 Theory and goals —— 1
1.1 Introduction —— **1**
1.2 Approach —— **1**
1.3 Terminology —— **3**
1.3.1 Metalanguage —— **3**
1.3.2 Terms —— **3**
1.3.2.1 Maltese —— **3**
1.3.2.2 Syntax —— **4**
1.3.2.3 Sentence —— **5**
1.3.2.4 Word —— **6**
1.3.2.5 Predicate —— **6**
1.3.2.6 Clause —— **7**
1.3.2.7 Phrase / Catena —— **7**
1.3.2.8 Constituent and constituent order —— **8**
1.3.2.9 Pragmatics —— **8**
1.4 Conclusion —— **8**

2 Approaches to the study of constituent order —— 11
2.1 Introduction —— **11**
2.2 The typological approach —— **12**
2.2.1 Greenbergian typology —— **12**
2.2.1.1 Greenberg 1963 —— **12**
2.2.1.2 Criticism and revision of Greenberg —— **13**
2.2.1.3 Evaluation —— **15**
2.3 The generative approach —— **15**
2.3.1 *Aspects* to Minimalism —— **15**
2.3.2 The Cartographic Project —— **17**

2.3.3 Constituent order and pragmatics in generativist thought — **19**
2.3.4 Discourse configurationality — **21**
2.3.5 Evaluation — **22**
2.4 Pragmatic approaches to constituent order — **23**
2.4.1 Functional Sentence Perspective (FSP) — **23**
2.4.2 Functional Generative Description (FGP) — **24**
2.4.3 Evaluation — **25**
2.4.4 Categorical and thetic judgments — **26**
2.5 The quantitative approach — **28**
2.6 Summary — **29**

3 Maltese constituent order: state of the question — 31
3.1 Introduction — **31**
3.1.1 General — **31**
3.2 Vella 1831 — **31**
3.2.1 Overview — **31**
3.2.2 Summary and evaluation — **32**
3.3 Sutcliffe 1936 — **33**
3.3.1 Overview — **33**
3.3.2 Summary and evaluation — **34**
3.4 Aquilina 1959 — **34**
3.4.1 Overview — **34**
3.4.2 Summary and evaluation — **35**
3.5 Vella 1970 — **36**
3.5.1 Overview — **36**
3.5.2 Summary and evaluation — **37**
3.6 Krier 1976 — **37**
3.6.1 Overview — **37**
3.6.2 Summary and evaluation — **39**
3.7 Kalmár and Agius 1983 — **39**
3.7.1 Overview — **39**
3.7.2 Summary and evaluation — **41**
3.8 Fabri 1993 — **43**
3.8.1 Overview — **43**
3.8.2 Summary and evaluation — **44**
3.9 Borg and Azzopardi-Alexander 1997 — **44**
3.9.1 Overview — **44**
3.9.2 Summary and evaluation — **45**
3.10 Fabri and Borg 2002 — **46**
3.10.1 Overview — **46**
3.10.2 Summary and evaluation — **47**
3.11 Other — **47**

3.12 Conclusion — **48**

4 Interlude: Research questions — 51
4.1 Introduction — **51**
4.2 Research questions — **51**
4.2.1 Research Question 1: What is the dominant constituent order in Maltese? — **51**
4.2.1.1 Question in context — **51**
4.2.1.2 How to answer it? — **52**
4.2.2 Research Question 2: What is the variation in dominant constituent order in Maltese? — **53**
4.2.2.1 Question in context — **53**
4.2.2.2 How to answer it? — **53**
4.2.3 Research Question 3: What are the deviations from the dominant constituent order in Maltese? — **53**
4.2.3.1 Question in context — **53**
4.2.3.2 How to answer it? — **53**
4.2.4 Research Question 4: What are the factors that cause variation in dominant constituent order? — **54**
4.2.4.1 Question in context — **54**
4.2.4.2 How to answer it? — **55**
4.3 Data and methodology — **55**
4.3.1 Data — **55**
4.3.2 Methodology — **56**

5 *BCv3*: A corpus of written Maltese — 57
5.1 Introduction — **57**
5.2 History of Maltese corpus linguistics — **57**
5.3 Data composition — **58**
5.3.1 Data selection — **58**
5.3.2 Text types — **59**
5.3.2.1 Text type: newspaper — **59**
5.3.2.2 Text type: parliament — **60**
5.3.2.3 Text type: fiction — **61**
5.3.2.4 Text type: non-fiction — **62**
5.3.2.5 Summary — **62**
5.3.3 Data processing — **63**
5.3.3.1 Text conversion and cleaning — **63**
5.3.3.2 Text cleaning — **63**
5.3.3.3 Sentence splitting — **63**
5.3.3.4 Tokenization — **64**
5.3.3.5 Language identification — **65**

5.3.3.6 Corpus management and querying — **65**
5.4 Enrichment — **66**
5.4.1 Part-of-speech tagging — **66**
5.4.1.1 The tagset — **66**
5.4.1.2 Tagging decisions and their hierarchy — **66**
5.4.1.3 Tags and their definition — **67**
5.4.1.4 Manual tagging — **80**
5.4.1.5 Automated tagging — **80**

6 Maltese Universal Dependencies Treebank v1 — 83
6.1 Introduction — **83**
6.2 Universal Dependencies — **83**
6.2.1 Why Universal Dependencies? — **83**
6.2.2 Levels of annotation and record format — **84**
6.3 Maltese UD annotation — **85**
6.3.1 ID — **85**
6.3.2 FORM — **85**
6.3.3 LEMMA — **85**
6.3.4 UPOSTAG: Universal part-of-speech tags — **85**
6.3.5 XPOSTAG: Maltese-specific part-of-speech tags — **86**
6.3.6 FEATS: Maltese morphological features — **86**
6.3.6.1 General — **86**
6.3.6.2 Lexical features: Poss — **87**
6.3.6.3 Nominal feature: Gender — **87**
6.3.6.4 Nominal feature: Animacy — **87**
6.3.6.5 Nominal feature: Number — **87**
6.3.6.6 Nominal feature: Case — **88**
6.3.6.7 Nominal feature: Definite — **88**
6.3.6.8 Nominal feature: Degree — **88**
6.3.6.9 Verbal feature: VerbForm — **88**
6.3.6.10 Verbal feature: Mood — **88**
6.3.6.11 Verbal feature: Tense — **89**
6.3.6.12 Verbal feature: Voice — **89**
6.3.6.13 Verbal feature: Person — **89**
6.3.6.14 Verbal feature: Negative — **89**
6.3.6.15 Note: Clitics — **89**
6.3.6.16 Morphological features: summary — **90**
6.3.7 HEAD: Head of the current word — **91**
6.3.8 DEPREL: Maltese universal dependency relations — **91**
6.3.9 DEPS: Enhanced dependency graph — **91**
6.3.10 MISC: Any other annotation — **91**
6.4 Maltese UD relations, or: a sketch of Maltese syntax — **91**

6.4.1 Introduction — **91**
6.4.2 General principles of syntactic annotation in UD v1 — **92**
6.4.3 Rules of syntactic annotation in UD v1 — **92**
6.4.4 Maltese UD relations — **95**
6.4.4.1 root — **95**
6.4.4.2 Core arguments: Valency frame — **112**
6.4.4.3 Core arguments: Nominals — **123**
6.4.4.4 Core arguments: Clauses — **128**
6.4.4.5 Non-core dependents: Nominals — **135**
6.4.4.6 Non-core dependents: Clauses — **140**
6.4.4.7 Non-core dependents: Modifier words — **141**
6.4.4.8 Non-core dependents: Function words — **143**
6.4.4.9 Nominal dependents: Nominals — **156**
6.4.4.10 Nominal dependents: Clauses — **158**
6.4.4.11 Nominal dependents: Modifier words — **159**
6.4.4.12 Nominal dependents: Function words — **160**
6.4.4.13 Coordination — **163**
6.4.4.14 Multi-word expressions — **165**
6.4.4.15 Loose relations — **168**
6.4.4.16 Special relations — **169**
6.4.4.17 Other relations — **171**
6.5 Data selection — **171**
6.5.1 Goals — **171**
6.5.2 Treebank composition — **172**
6.5.3 Manual annotation — **177**
6.5.4 Corpus management and querying — **181**

7 Dominant constituent order and its variations in Maltese: A quantitative analysis — 183
7.1 Introduction — **183**
7.2 Basic statistics — **183**
7.2.1 Sentence length and complexity — **183**
7.2.2 Clause types — **187**
7.2.2.1 General — **187**
7.2.2.2 UD clause types — **187**
7.2.2.3 Clause types by root — **188**
7.3 Constituent order in MUDTv1 by the numbers — **189**
7.3.1 Overview — **189**
7.3.2 Verbal clauses — **192**
7.3.2.1 Introductory remarks — **192**
7.3.2.2 Order of active subject and predicate — **192**
7.3.2.3 Order of predicate and direct object — **203**

7.3.2.4 Order of predicate and indirect object — **208**
7.3.2.5 Order of passive subject and predicate — **211**
7.3.2.6 Order of predicate and passive agent — **216**
7.3.3 Copular clauses — **218**
7.3.3.1 Overview — **218**
7.3.3.2 VS in copular xcomp — **220**
7.3.4 Existential clauses — **223**
7.3.5 Constituent order across text types — **225**
7.4 A brief comparison, or: a két fadatbázis regénye — **227**
7.4.1 Introduction — **227**
7.4.2 Data and analysis — **228**
7.4.3 Conclusion — **233**
7.5 Summary — **238**
7.5.1 Introduction — **238**
7.5.2 Answer to Research Question 1: What is the dominant constituent order in Maltese? — **238**
7.5.3 Answer to Research Question 2: What is the variation in dominant constituent order in Maltese? — **238**
7.5.4 Answer to Research Question 3: What are the deviations from the dominant constituent order in Maltese? — **239**
7.5.5 Answer to Research Question 4: What are the factors that cause variation in dominant constituent order? — **239**
7.5.6 Final considerations — **239**

8 Summary — 243
8.1 The road up here — **243**
8.2 The road ahead — **244**

Bibliography — 247

Abbreviations — 261

Appendix — 263

List of Tables

Tab. 5.1 Text types in *BCv3* —— **59**
Tab. 5.2 Text type newspaper in *BCv3* —— **60**
Tab. 5.3 Text type parliament in *BCv3* —— **61**
Tab. 5.4 Text type fiction in *BCv3* —— **61**
Tab. 5.5 Text type non-fiction in *BCv3* —— **62**
Tab. 5.6 Text types in *BCv3* —— **62**
Tab. 5.7 Maltese part-of-speech tagset —— **67**
Tab. 5.8 SVMTool part-of-speech tagging accuracy —— **81**

Tab. 6.1 Levels of annotation in UD / CoNLL-U format —— **84**
Tab. 6.2 Universal part-of-speech tags —— **85**
Tab. 6.3 UD v1 morphological features —— **86**
Tab. 6.4 UD v1 morphological features in Maltese —— **90**
Tab. 6.5 UD v1 relations adapted to and extended for Maltese —— **94**
Tab. 6.6 Actants (core dependents) in VALLEX —— **116**
Tab. 6.7 Free dependents in VALLEX —— **117**
Tab. 6.8 MUDTv1 composition: summary —— **174**
Tab. 6.9 MUDTv1 composition: text types newspaper and parliament —— **175**
Tab. 6.10 MUDTv1 composition: text types fiction and non-fiction —— **176**

Tab. 7.1 MUDTv1: Mean sentence length by text type —— **184**
Tab. 7.2 MUDTv1: Mean sentence length in written and quasi-spoken texts —— **185**
Tab. 7.3 MUDTv1: UD clause types —— **188**
Tab. 7.4 MUDTv1: Clauses containing core dependents by root (columns) and UD clause type (rows) —— **189**
Tab. 7.5 MUDTv1: Constituent order – Greenbergian classification —— **190**
Tab. 7.6 MUDTv1: Order of subject and predicate in active clauses by UD clause type —— **194**
Tab. 7.7 MUDTv1: Dominant VS in active xcomp —— **194**
Tab. 7.8 MUDTv1: Head of active xcomp —— **196**
Tab. 7.9 MUDTv1: Dominant VS in active parataxis —— **197**
Tab. 7.10 MUDTv1: Dominant VS in active parataxis: Types —— **197**
Tab. 7.11 MUDTv1: Classification of constituent order variation in active parataxis —— **198**
Tab. 7.12 MUDTv1: Order of S and V in active acl —— **199**
Tab. 7.13 MUDTv1: Order of predicate and direct object in verbal clauses by UD clause type —— **205**
Tab. 7.14 MUDTv1: Types of OV clauses —— **206**
Tab. 7.15 MUDTv1: dobj and nmod:obj —— **208**
Tab. 7.16 MUDTv1: Order of predicate and indirect object in verbal clauses by UD clause type —— **210**
Tab. 7.17 MUDTv1: Order of predicate, direct object and indirect object —— **211**
Tab. 7.18 MUDTv1: Order of subject and predicate in passive clauses by UD clause type —— **213**
Tab. 7.19 MUDTv1: 33% threshold for determining dominant constituent order in borderline cases —— **214**
Tab. 7.20 MUDTv1: Dominant VS in passive xcomp —— **216**

Tab. 7.21 MUDTv1: Head of passive xcomp —— **216**
Tab. 7.22 MUDTv1: Order of passive agent and predicate in passive clauses by UD clause type —— **218**
Tab. 7.23 MUDTv1: Order of subject and predicate in copular clauses by UD clause type —— **220**
Tab. 7.24 MUDTv1: Dominant VS in copular xcomp —— **220**
Tab. 7.25 MUDTv1: Classification of constituent order variation in copular xcomp —— **223**
Tab. 7.26 MUDTv1: Order of subject and predicate in existential clauses by UD clause type —— **225**
Tab. 7.27 MUDTv1: Ratio of VS and OV across text types —— **226**
Tab. 7.28 MUDTv1: Variation in dominant constituent order —— **238**

List of Figures

Fig. 6.1 PosTagger: Syntactic annotation —— **178**
Fig. 6.2 PosTagger: Relation selection —— **179**

Fig. 7.1 MUDTv1: Mean sentence length by file and text type —— **184**
Fig. 7.2 MUDTv1: Mean sentence length in written and quasi-spoken texts —— **185**
Fig. 7.3 MUDTv1: Sentence complexity by text type —— **186**
Fig. 7.4 MUDTv1: Sentence complexity by text type —— **187**
Fig. 7.5 MUDTv1: Constituent order – Greenbergian classification —— **190**
Fig. 7.6 MUDTv1: Constituent order by clause type – overview —— **191**
Fig. 7.7 MUDTv1: Order of subject and predicate in active clauses by UD clause type —— **193**
Fig. 7.8 MUDTv1: Subject heaviness and clause length as predictors of the order of S and V in active acl clauses —— **203**
Fig. 7.9 MUDTv1: Order of predicate and direct object in verbal clauses by UD clause type —— **204**
Fig. 7.10 MUDTv1: Order of predicate and indirect object in verbal clauses by UD clause type —— **209**
Fig. 7.11 MUDTv1: Order of predicate, direct object and indirect object —— **211**
Fig. 7.12 MUDTv1: Order of subject and predicate in passive clauses by UD clause type —— **212**
Fig. 7.13 MUDTv1: Sampled VS order in borderline cases with main as control —— **215**
Fig. 7.14 MUDTv1: Order of passive agent and predicate in passive clauses by UD clause type —— **217**
Fig. 7.15 MUDTv1: Order of subject and predicate in copular clauses by UD clause type —— **219**
Fig. 7.16 MUDTv1: Order of subject and predicate in existential clauses by UD clause type —— **224**
Fig. 7.17 MUDTv1: Order of subject and predicate by text type and UD clause type —— **226**
Fig. 7.18 MUDTv1 vs HUUDv2: Constituent order – Greenbergian classification —— **229**
Fig. 7.19 MUDTv1 vs HUUDv2: Order of subject and predicate by UD clause type —— **230**
Fig. 7.20 MUDTv1 vs HUUDv2: SV vs VS order —— **231**
Fig. 7.21 MUDTv1 vs HUUDv2: Order of predicate and object by UD clause type —— **232**
Fig. 7.22 MUDTv1 vs HUUDv2: VO vs OV order —— **233**
Fig. 7.23 MUDTv1 vs HUUDv2: VO vs OV order with 100 additional OV clauses in MUDTv1 per clause type —— **234**
Fig. 7.24 Greenbergian classification of Maltese, Hungarian, Greek and English —— **236**
Fig. 7.25 Dryerian classification of Maltese, Hungarian, Greek and English —— **237**

1 Theory and goals

1.1 Introduction

This thesis is, as apparent from the title, a study of constituent order in Maltese and as such, it is a work on Maltese syntax and (to a smaller extent) pragmatics. The first chapter of a thesis is normally the place for setting the research questions and describing the data and methodology employed in answering them and I will get to that in due course. Before I do, however, there are broader issues to be discussed, issues of fundamental importance that are often taken for granted or downright ignored, like the nature of linguistics, its goals and its methods. One of the greater insights I have gained in the work on this dissertation is that the failure to consider these issues seriously undermines the scientific enterprise. In this chapter, I will therefore provide answers to these questions not only to avoid the pitfalls described above, but also to make it clear what this thesis is and what it is not.

1.2 Approach

The general approach I employ in this thesis is best described using the adjectives "descriptive" and "empirical". What follows is the definition of those terms and the reasoning behind them.

It is my view that the primary task of linguistics is to describe a language i.e., in Haspelmath's (2009: 344) definition, to provide a "characterization of grammatical regularities" of a language. In much of linguistic literature, the descriptive approach is contrasted with the theoretical approach, where the latter is rooted in a particular framework, i.e. "a sophisticated and complex metalanguage for linguistic description intended to work for any language" (Haspelmath 2009: 343). To pick a random example from my library, Lieber and Štekauer in their introduction to a handbook of compounding (Lieber and Štekauer 2009: 3-4) speak of complicating the view of the subject of their research "both theoretically and descriptively" where the former involves "consider[ing] compounding from disparate frameworks" and the latter entails "looking not only at familiar languages, but also at a range of typologically and areally diverse languages". It is primarily within the context of that dichotomy that I wish to characterize my approach as descriptive and framework-free: I aim to provide a description of a part of the grammar of a particular language while doing so outside of any existing theoretical framework, i.e. considering the language on its own, without any conscious preconceptions or biases.

In this sense, my "description" is essentially equivalent to Haspelmath's (2004) "phenomenological description", but perhaps narrower: Haspelmath argues that phenomenological description entails accurate prediction of speaker behavior (Haspel-

math 2009: 344). I find this framing troubling for several reasons (such as the ultimate utility of prediction with regard to such a complex and downright chaotic system as a human being or the use of the term "behavior" in reference to language) and therefore prefer to speak of a description that accounts for the data. The traditional adjective applied in these circumstances is empirical. I will gladly accept it in the context of its contrast to introspective or intuitive approaches (see Itkonen 2005). The corporate world, which I used to inhabit, typically describes my approach as data-driven, a label that strikes me as even more apt and still relatively baggage-free.

The dichotomy between descriptive and theoretical linguistics goes deeper than the choice of metalanguage: as Haspelmath (2004: 555) notes, some linguists view their primary task to be not linguistic description as defined above, but either (i) the creation of an accurate representation of the speakers' mental grammars or (ii) the description of the "cognitive code" for language (also known as "faculty of language", "Universal Grammar", "I-language" or just "Language"). One's view of the attainability of these specific goals largely depends on whether one is convinced of the existence of these phenomena, but even if one is, they can only be studied by inference. In this modern rehash of the universals debate (cf. Katz 1996), I will confess myself to be a nominalist: only language as spoken and written by its speakers ("parole", "performance" or simply "language") is a real thing that exists here and now and that can be observed directly. Unlike my realist colleagues and other critics of linguistic nominalism,[1] I view the study of both "Language" as well as "language" as equally worthy endeavors and so in my view, the choice depends on one's preferences and priorities. Mine should be obvious from the subtitle of this thesis.

And this brings me to my final point: it should be noted that while corpus linguistics is often seen as the pinnacle of empiricism in linguistics, it is not a synonym for it. Experience has shown that corpus data, however large, may not be sufficient for a complete description of some linguistic phenomena such as complex morphology (Cvrček et al. 2015: 22) or low-frequency syntactic constructions (Pullum 2017). Any full description of any language should thus make full advantage of all data collection tools available to a linguist, including elicitation and experimentation. This work, being the first detailed treatment of its titular subject and a doctoral thesis, is merely the first step towards the

1 One would be tempted to cite here Chomsky's famous "butterfly collecting" remark (Chomsky and Ronat 1979: 57), but why beat a dead horse. Instead, I will direct the readers' attention to the following comment made by Norbert Hornstein of University of Maryland, a prominent generativist, who perfectly summarizes the views of a large portion of Chomsky's modern followers on the subject: "I don't much care about language. I care about FL (Faculty of Language, ed.) and it's (sic) structure. That's what GG studies. That's what I find interesting." (bit.ly/2FEveam, last consulted on February 28 2018). Even Katz's realist rejection of generative grammar (1996) argues that "grammars should be seen as scientific theories, not as data-cataloguing devices" (Katz 1996: 292) and, much like the object of his criticism, rejects empiricism by insisting that "Grammatical questions are factual questions, but they are no more empirical questions than mathematical or logical questions" (Katz 1996: 292).

full description of the phenomenon and as such, its goals and the tools used to achieve them are deliberately narrow.

1.3 Terminology

1.3.1 Metalanguage

While this work eschews the use of any particular framework and strives to describe its object of study on its own terms, a metalanguage is nevertheless necessary for the description of the phenomena it sets out to study. An elegant and clean solution to this conundrum would be to come up with a completely new one, free from existing Latin- or English-based biases inherent to concepts such as "noun" or "verbal phrase". This, much like most elegant and clean solutions, is neither practical nor reasonable, and so in what follows, I will employ a compromise and use terms that are largely familiar to anyone who has ever read a grammar, but with their meaning extended or narrowed as necessary. This is a solution that has been practiced by linguists for centuries – just consider the different meanings of the terms "imperfect" and "perfect" when referring to the verbal system of Latin and when applied to the verbal system of Semitic languages (going at least as far back as Wahrmund 1861). Its only downside is that in the olden days, there were only a few sources of linguistic metalanguage, whereas by now our field has accumulated so much terminology that confusion is difficult to avoid. In my work, I often find myself using two terms synonymously even though to their originators and other users, these might have two very different referents. I will therefore use this section to provide definitions of the fundamental terms I will use and the concepts behind them.

1.3.2 Terms

1.3.2.1 Maltese

This thesis aims to study Maltese (endonymic glossonym *l-ilsien Malti*, *il-lingwa Maltija* or simply *il-Malti*) which in general terms refers to the Semitic language descended from North African dialects of Arabic (Jastrow 1980: 286-291, Corriente and Ángeles 2008: 379-407) spoken primarily in the Maltese archipelago (352,121 speakers)[2] and Australia (34,396 speakers).[3]

2 This figure includes only those respondents in the 2011 census aged 10 or older who reported speaking Maltese "Well", excluding the 5,571 who describe their competence in Maltese as "Average" and the 8,174 who reported speaking "A little" Maltese. For details, see p. 149 of bit.ly/2thUvng (last consulted on February 28th 2018).

3 2011 census data, see bit.ly/2yO3HCo (last consulted on February 28th 2018).

However, for the purposes of the analysis employed herein, "Maltese" shall refer solely to the written language produced by native speakers of Maltese in the first two decades of the 21st century as represented by the texts contained in the two corpora used as the source of data for the analysis herein (see Chapter 5 and 6). The question of to what extent these are representative of the language as used by her speakers is a complex one, deserving of a more detailed treatment than can be given here. It will therefore be addressed in the relevant parts of this thesis only to the extent that is necessary for the complete description of the data set. Additionally, some of the texts in the two corpora originated as spoken language (e.g. journalistic interviews or Parliament speeches), yet the process by which they were turned into written text has been found to be unreliable and even to distort the original. I will therefore treat those texts as if they originated in writing, albeit as a distinct text type (see Chapter 6, section 6.5.2), and will refrain from making any judgments on the spoken language that underlies them.

1.3.2.2 Syntax

As noted above, this is a work on syntax. In what follows, I will frame the discussion of syntax from the point of view of dependency syntax. This may seem like a contradiction given my insistence on the framework-free nature of my approach to studying syntax, but it is not: as you will recall, I defined a linguistic framework as "a sophisticated and complex metalanguage for linguistic description intended to work for any language" (Haspelmath 2009: 343). Dependency syntax is neither sophisticated,[4] nor complex, in fact, it isn't even a metalanguage; it is merely one particular way of thinking about and formalizing syntax, ultimately reducible to a handful of fundamental principles (*Elements*, Chapter 1 through 3; Tesnière 1959: 11-15, Tesnière 2015: 3-6):[5]

1. The object of syntax is to study the sentence in terms of the relationships between its words (*Elements*, Chapter 1, §1-11).
2. Those relationships are defined in terms of a governor and a subordinate or dependent (*Elements*, Chapter 2, §1-3).
3. The subordinate typically depends on one and exactly one governor (Elements, Chapter 3, §1).
4. There exists a hierarchy of dependencies with one word ultimately governing all the rest (*Elements*, Chapter 3, §5-6). This word (the central node or root) is typically, but not always, a verb (*Elements*, Chapter 3, §7).

4 This is a descriptive statement, not an evaluative one.

5 Throughout this work, each reference to *Éléments de syntaxe structurale* will be accompanied by the references to the original text and the 2015 English translation only on the first occurrence in a chapter. Subsequent references will contain the shortened name of the work as above, chapter number and – wherever applicable – the section (§) number.

There are a number of approaches to dependency syntax, from the relatively simple Stanford Dependencies (de Marneffe and Manning 2008) to the elaborate annotation of semantic relationships, discourse relations, anaphora relationships and multiword expressions in the Prague Dependency Treebank 3.0 (Bejček et al. 2013); there even exist a number of theoretical frameworks based on the principles of dependency linguistics (Ágel et al. 2003: 508-716). When defined in opposition, dependency linguistics should thus be contrasted with the general category of phrase structure grammars (Matthews 1981: 71-72), rather than a specific framework.

The particular approach I've chosen here is that implemented by the Universal Dependencies project (henceforth: UD, de Marneffe, Dozat et al. 2014). This flavor of dependency linguistics is the very definition of a framework-free approach to linguistic description and comparison as per Haspelmath (2009): it is a set of labels for comparative concepts (in this case syntactic relationships) which will not fit every language perfectly, but can be stretched or shrunk as required by the facts of a particular language. Haspelmath (2009: 363) is quick to note that such "comparative concepts are not necessarily equatable with the descriptive categories of languages", but makes it clear that they can be. To show that, he cites an example from *The World Atlas of Language Structures* (WALS) where the term "case" is used differently by different authors (see Iggesen 2013) noting that "The concepts are not identical, only the chosen terms happen to coincide" (Haspelmath 2009: 364). If the transfer goes one way, it goes the other as well (see also the discussion of portable and non-portable labels in Haspelmath 2017). In my analysis, I have therefore opted to apply this principle to the syntactic analysis of Maltese by using the concepts of UD and widening or narrowing them as necessary. Chapter 6 discusses this adaptation of UD to Maltese in detail.

One last note: throughout this work (especially when discussing annotation decisions), I will use the adjective "syntactic", as in "syntactic criteria" or "syntactic role". This should be taken to mean that in my analysis (e.g. of what part of speech to assign to a particular word or what label to use for a particular construction), I am not guided by morphology or semantics, but rather by what the governor-dependent relationships in the particular phrase or clause are.

1.3.2.3 Sentence

As noted above, the object of syntax is the study of the sentence. The definition of what a sentence is, however, is far from clear-cut and definitely not generally agreed upon. One attempt to provide an answer to this question is famous for its comprehensive list of definitions (Ries 1931: 208-224). In the face of this, some overviews of syntax start their discussion of what a sentence is by citing dictionary definitions (Matthews 1981: 26, Dürscheid 2007: 58), a sure sign of desperation.

As this work focuses on written language, many of the issues faced by those who attempt to define a "sentence" can be avoided, chief among them the issue of the existence of sentence in spoken language (Miller 1995, Halliday 2014: 428-438). The pri-

mary problem to be resolved here is that of the status of what Culicover and Jackendoff (2005: 236-238) refer to as "nonsentential utterance types", such as single-words, vocatives, expletives and even items like numbers and list delimiters. While these present serious challenges to some types of linguistic frameworks (such as that proposed by Culicover and Jackendoff 2005), in dependency linguistics, this is merely another instance of determining the root and its dependents.

In this thesis, the term "sentence" is therefore defined in technical and practical terms as "orthographic sentence resulting from the process of sentence splitting as described in Chapter 5, section 5.3.3.3", regardless of the number of words (see below) and its structure.

1.3.2.4 Word

Much like "sentence", "word" is another problematic concept with definitional issues ranging from phonological through morphological all the way to syntactic and orthographic (see Dixon and Aikhenvald 2003 for a thorough analysis). The definition of what a word is is especially relevant for the type of analysis conducted here since, as noted above, dependency syntax is concerned with the relationship of words within a sentence. This is, naturally, true of UD as well, except UD does not provide a standard definition of a word; this decision is left to the creators of individual treebanks based on the requirements of their respective languages. The only guideline the UD v1 standard provides is that "the basic units of annotation are *syntactic* words" (Nivre, Ginter et al. 2014, italics in the original) and not orthographic words. This has consequences for languages that employ contractions like English and French or those that attach clitics to their orthographic words like Spanish, Italian. Maltese also belongs to the latter group, however, due to complexities of its morphology (discussed in Chapter 5, section 5.3.3.4), the issue of defining a syntactic word is a complex one. For the purposes of this thesis, the term "word" is therefore defined in technical and practical terms as "token resulting from the process of tokenization as described in Chapter 5, section 5.3.3.4."

In this context, and line with common corpus linguistics usage, the term "type" when used in conjunction with the term "token" will refer to unique forms of which a particular token is an instantiation (McEnery and Hardie 2012: 50).

1.3.2.5 Predicate

Dependency linguistics can be considered the product of the rejection of the traditional division of sentence into subject and predicate (ultimately traceable to Aristotle, Graffi 2001: 75) and its substitution with the principle of verb-centrality (*Elements*, Chapter 48-49). As Tesnière himself notes, however, "nothing prevents a sentence from having a noun as its central node, or an adjective or an adverb" (*Elements*, Chapter 3, §7). This is doubly true of languages which lack an overt copula or only use it in some contexts, such as Russian, Hungarian and Maltese, and so make use of non-verbal sentences. For this reason (and in line with the preferred usage of the UD standard; Nivre, Ginter et al.

2016: 1661), the term "predicate" will be used here for "the central node or root of a sentence or clause", be it a verb or any other part of speech.

1.3.2.6 Clause

A sentence may contain several predicates which are, in accordance with the principles of dependency linguistics, in a governor-dependent relationship to one another (whether directly or through one of its dependents) with one predicate ultimately dominating all the others (again, whether directly or through one of its dependents). Each single predicate with all of its dependents will be referred to as a "clause". The clause dominating all the others will be referred to as the "main clause", all other clauses will be – solely for convenience – generically referred to as "dependent clauses" regardless of their actual syntactic relationship to the main clause. A detailed classification of dependent clauses in terms of UD is discussed in Chapter 6.

1.3.2.7 Phrase / Catena

While I frame the description of sentence syntax in terms of dependency syntax, I will nevertheless occasionally use the term "phrase", mostly "noun phrase" or "prepositional phrase". These terms must then be taken to mean either "a group of words dominated by a single noun or pronoun" and "a group of words dominated by a single noun or pronoun which governs at least one preposition", respectively. This is one of those instances where I use a technical term specific to a particular school of thought that I have repurposed for a concept completely at odds with its original meaning, even sacrificing consistency. I did so for reasons of familiarity and brevity, as is evident from the definition of "prepositional phrase" which in its original definition is contrary to the lexicalist principles of UD (see Chapter 6, section 6.4.2), but is nevertheless preferable to the cumbersome formulation above.

The term "catena" used in some dependency grammars (Osborne et al. 2012) also satisfies the brevity criterion and could therefore be employed here, even if its relatively low currency speaks against it. I may nevertheless use it whenever appropriate (especially in descriptions of dependency graphs) and in such cases, it should be read as synonymous with "phrase".

In this context, the reader will also encounter the terms "structure" and "construction", especially in discussion of various examples illustrating dependency relations. These are used generically (as opposed to the framework-specific term "syntactic structure") to describe a particular configuration of governor and its dependents to express a particular set of grammatical relations; as such, they should be taken to be synonymous with "phrase".

1.3.2.8 Constituent and constituent order

In what follows, I will maintain a strict distinction between constituent order and word order. This is another contradiction to the dependency-based analysis employed here, as dependency grammars deal with relationships between words and thus anything that involves the order and the sentence is, by definition, word order. I have established the distinction for reasons of clarity: this thesis is only concerned with the order in which the predicate and its core arguments (as defined by UD v1, see Chapter 6, section 6.4.3) appear in a sentence. This will be termed "constituent order". In contrast, the order of elements with a phrase (such as the order of nouns and adjectives or adjectives and adverbs) will be referred to as "word order" and will not be addressed here, save in passing. All of this, of course, applies only to my own words; in citations from other works, I will use terminology employed by the respective author.

In this context, I will use the term "configuration" in the sense of "possible arrangements of the predicate and its dependents". The term "configurationality", in contrast, will only be used in reference to the generativist concept (cf. Chapter 2, sections 2.3.2 and 2.3.3) and those works or authors who employ it.

Chapter 2 discusses various theories of constituent order in detail.

1.3.2.9 Pragmatics

In this thesis, the term "pragmatics" will be used in its broad sense to mean – borrowing a definition from Levinson's classic textbook (Levinson 1983: 27) – "the study of deixis (at least in part), implicature, presupposition, speech acts, and aspects of discourse structure." In other words, while syntax studies the structure of sentences, pragmatics studies the context in which they are uttered/written and how it influences their structure.

Information structure is a subfield of pragmatics, ultimately traceable at least as far back as Mathesius's functional division of a sentence juxtaposed with its formal division (Mathesius 1939). In its classic definition, it is the "structuring of sentences by syntactic, prosodic, or morphological means that arises from the need to meet the communicative demands of a particular context or discourse" (Vallduví and Engdahl 1996: 460).

1.4 Conclusion

In what follows, I will apply the principles and the conceptual apparatus outlined above to the study of constituent order in Maltese. The first steps in that process involve surveying existing literature in order to show why studying constituent order is important, how to do it and how not to, what has already been done in that regard for Maltese and what is there still to do. Chapters 2 and 3 are devoted to that purpose and thus to setting

stage for defining the research questions and the way in which I will go about answering them; Chapter 4 will then be the place for that.

2 Approaches to the study of constituent order

2.1 Introduction

Constituent order is one of the fundamental elements of syntactic description. Its importance is evidenced by the fact that it is often the only piece of information available on the syntax of a language; indeed as Dixon (2013: 73) notes, since the most of the world's languages are under-described, it is often the only piece of information on the grammar of a language available. Comprehensive overviews of the world's languages such as *Ethnologue* (Lewis et al. 2016) are the best witness to this. To pick two random examples, the *Ethnologue* entry for Swedish (ISO 639-3 code "swe"), a relatively small but well-described language, lists the following under "Typology":

> SVO; prepositions; noun head final; gender (common, neuter); definite and indefinite articles; passives (active, middle, passive); comparatives; 19 consonant and 17 vowel phonemes; tonal (2 tones).

For Övdalian (ISO 639-3 code "ovd"), also spoken in Sweden, a close relative of Swedish and thus hardly an exotic language, the same section contains only the following:

> SVO; 24 consonants, 9 vowels, 6 diphthongs and 1 triphthong.

The noticeably frequent appearance of constituent order in even the most rudimentary language descriptions is likely due to two factors: first, constituent order is typologically associated with a number of other syntactic and even morphological features (see section 2 below) and can thus serve as a microcosm of a language's grammar. Secondly, constituent order is one of those properties of a language that are conspicuous (especially when different from what one is used to) and thus seem relatively easily discernible, much like its phonological inventory (again, see the Övdalian example above). As such, constituent order attracted the attention of linguists even in times when analyses of syntax rarely went beyond rudimentary descriptions of simple and compound sentences (Weil 1844)[1] and with the advent of modern syntactic theories, the literature grew exponentially.

In this chapter, I provide an overview of approaches to the study of constituent order current in modern linguistics, in rough order of their prevalence or popularity. This overview is not intended to be exhaustive or comprehensive (such a task is well beyond the scope of this work) and may include more obscure schools of thought. The primary purpose of this chapter is to inform the following discussion, especially that

1 Though to be exact, his and similar works (see section 2.4.4 below) had as much to do with the conflation of grammatical and logical categories of subject and predicate in Western philosophical tradition as with constituent order variation.

of previous works on constituent order in Maltese (Chapter 3) and the formulation of the research questions. The selection of the approaches discussed here is tailored specifically to that purpose.

2.2 The typological approach

2.2.1 Greenbergian typology

2.2.1.1 Greenberg 1963

The undoubtedly most influential work on constituent order in modern linguistics is Joseph H. Greenberg's 1963 paper titled *Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements* (cited from the second edition, Greenberg 1966). Greenberg expanded relatively trivial observations on how languages differ in the order of "modifying or limiting elements" (Greenberg 1966: 76) into a full-fledged typological classification of languages based on a list of so-called universals. The fundament on which these rest is his basic order typology: Greenberg takes the observation that "languages have several variant orders but a single dominant one" (Greenberg 1966: 76) to its logical conclusion and establishes a six-way typology of dominant orders of subject, verb and object: SVO, SOV, VSO, VOS, OSV and OVS. He immediately notes, however, that three of those – VOS, OSV and OVS – "do not occur at all, or at least are rare" (Greenberg 1966: 76) and proceeds to draw from this his first universal:

> *Universal 1*. In declarative sentences with nominal subject and object, the dominant order is almost always one in which the subject precedes the object.

Greenberg combines the remaining three configurations – referred to as Type I (VSO), Type II (SVO) and Type III (SOV) – with two additional binary criteria (whether a language has prepositions or postpositions and whether an adjective of quality follows the noun it modifies or precedes it) and investigates the correlations between these syntactic properties in a sample of 30 languages (Greenberg 1966: 74-75):

> Basque, Serbian, Welsh, Norwegian, Modern Greek, Italian, Finnish (European); Yoruba, Nubian, Swahili, Fulani. Masai, Songhai, Berber (African); Turkish, Hebrew, Burushaski, Hindi, Kannada, Japanese, Thai, Burmese, Malay (Asian); Maori, Loritja (Oceanian); Maya Zapotec, Quechua, Chibcha, Guarani (American Indian).

Using these correlations as the starting point, Greenberg postulates 45 implicational universals, 15 of which relate to constituent order or at least the position of the verb and its arguments, including question words.

2.2.1.2 Criticism and revision of Greenberg

As with any dominant paradigm, criticisms of Greenberg began to appear almost immediately (Vennemann 1974 and 1976, Lehmann 1978, Hawkins 1983, to name but a few prominent ones). Vennemann's critique seeks to integrate Greenbergian universals with historical linguistics ("[t]he present paper has been concerned with causes of syntactic change", Vennemann 1974: 370), as well as information structure concerns (cf. the very title of Vennemann 1974). He arrives at an alternative typology consisting of two types binary oppositions: VX/XV (VO/OV) and operator/operand (cf. Tesnière's governor/dependent). Vennemann's typology is built around the ordering of operators and operands, a process to which Vennemann refers as "natural serialization" (cf. Tesnière's "linearization"; *Elements*, Chapter 14; Tesnière 1953: 32-33, Tesnière 2015: 25-26): according to his intepretation of the available data, VX is associated with operand-operator order, while operator-operand is characteristic of XV languages. Needless to say, these concepts are identical to Tesnière's division of languages into "centrifugal" and "centripetal" (*Elements*, Chapter 14) and also to head-initiality/head-finality of (post-)Chomskyan linguistics.

One of the primary issues that emerged as a major point of contention is the problem of basic (default) word order. Greenberg's original formulation of his universal doesn't actually define what qualifies as "basic", merely assumes it: "If a language has verb-subject-object as its basic word order in main declarative clauses..." (Greenberg 1966: 74). Greenberg is aware that this presupposes, at the very least, the existence of a subject-predicate structure in all languages under investigation. He acknowledges the problems with this assumption, but proceeds without resolving this issue since doing so would have "prevented me from going forward to those specific hypotheses, based on such investigation, which have empirical import and are of primary interest to the nonlinguist" (Greenberg 1966: 74).

Immediate responses to Greenberg like Vennemann (1974: 344) still rely on Greenberg's (lack of) definition, but others note their inherent problematic nature. Hawkins' discussion of the concept of basic order (1983: 12-13) addresses the primary issue, the existence of two competing orders (see also Comrie 1989: 88-89 and his discuss of "split order") in some languages and/or some types of structures, citing such notorious examples as the SOV order in German subordinate clauses. In order to resolve the conundrum, he establishes " three (overlapping) criteria when making a basicness decision" (Hawkins 1983: 13): frequency in attested samples, frequency in the grammatical system and lack of markedness (whether morphological or syntactic). This introduces a near constant theme in the definition of basic order, the conflict between (lack of) markedness (or neutrality, however defined) and frequency as the primary criterion in determining the basic order.

Matthew Dryer stands out from among the great number of Greenberg's successors as one of his most important critics. Dryer's work on constituent order typology began as a criticism of Greenberg's sampling methods and a test of hypotheses raised

by the Greenbergian universals (Dryer 1989b) and included a large follow-up study of the universals using a larger and more balanced sample of languages (Dryer 1992). In addition to this, Dryer (1992) offers a partial critique of Vennemann's work: he accepts the VO/OV typology, but argues against the operator-operand part of it (which he, in a fit of Rossianism, refers to as the "Head-Dependent Theory" or HDT) and instead proposes a replacement, the Branching Direction Theory (BDT), "according to which the word order correlations reflect a tendency for phrasal categories to precede nonphrasal categories in OV languages and vice versa in VO languages" (Dryer 1992: 132-133). Dryer thus introduces a new version of the concept of "right-branching" and "left-branching" into the typology of constituent and word order.

All this lead Dryer to renounce the Greenbergian six-way typology and propose a new typology, based on two independent but interacting binary parameters, SV/VS and VO/OV (Dryer 1997, Dryer 2013b). Dryer lays out a complex case for this, the chief arguments being that "some word order parameters correlate with both the order of the object and the verb and with the order of the subject and the verb" (Dryer 2013b: 295) and that a typology based on these two parameters is more fundamental than the six-way typology, as it is "based on clause types that occur much more frequently" (Dryer 1997: 70). The latter illustrates Dryer's focus on frequency as an important element in linguistic description and explanation: Dryer recognizes that "speakers store grammatical knowledge independent of frequency", but argues that "frequency plays a pervasive role in explaining why languages — and grammars — are the way they are" (Dryer 2013b: 292). Consequently, Dryer's concept of basic order is based on frequency where, admirably, Dryer is aware of the inherent dangers of inadequate sampling (Dryer 1997: 72, italics in the original):

> If a particular order is more common in most or all texts, then we can justifiably describe that order as most frequent. If no order is most frequent over most texts, however, or if the order varies from genre to genre or text to text, we should probably not describe any particular order as the basic order (in the sense of most frequent order) and we should say that the language is one that lacks a basic word order [...]. In short, while it may be relatively easy to identify a most frequent order in a single text or in a small body of texts, it is necessary to examine a wide variety of texts before one can decide with confidence that a particular order is most frequent in the language *as a whole*.

In typological studies of word and constituent order, Dryer's work has become the standard reference, as evidenced not only by his contribution to general discussions on the state of the question (see the special issue of *Linguistic Typology* 15), but also his authorship of chapters on word order in such overview of language typology as Shoppen 2007 (Dryer 2007) *The World Atlas of Language Structures* (WALS, Dryer 2013a and 2013c). And while the latter work uses the Greenbergian six-way typology in its description of constituent order typology (undoubtedly an editorial compromise), it is here that he provides the ultimate definition of basic order defined in terms of frequency (Dryer 2013a):

> The expression dominant order is used here, rather than the more common expression basic order, to emphasize that priority is given here to the criterion of what is more frequent in language use, as reflected in texts. ... The rule of thumb employed is that if text counts reveal one order of a pair of elements to be more than twice as common as the other order, then that order is considered dominant, while if the frequency of the two orders is such that the more frequent order is less than twice as common as the other, the language is treated as lacking a dominant order for that pair of elements. For sets of three elements, one order is considered dominant if text counts reveal it to be more than twice as common as the next most frequent order; if no order has this property, then the language is treated as lacking a dominant order for that set of elements.

This definition (applicable to both word order and constituent order, and both pairs and triads) is specific, empirically founded, without any theoretical baggage, cross-linguistically applicable and clearly actionable (step 1: get texts; step 2: count); as such, it constitutes a significant improvement to previous definitions of "basic" word (and constituent) order.

2.2.1.3 Evaluation

Greenberg's universals were met with almost immediate acceptance and despite substantial criticism (on which see above) and some empirical evidence to the contrary (like the case of OVS order in Hixkaranya described by Derbyshire 1977), Greenberg's six-way typology continues to be the dominant paradigm in the cross-linguistic study of constituent-order variation. Works like Payne (1997: 71-74), Song (2011b), the *Ethnologue* (see the entries above) and WALS are but a few of the most prominent examples of Greenbergs's enduring legacy.

In operational terms, Greenberg and the many follow-up studies to and refinements of his work and theory have provided a solid background for the study of inter-language constituent order variation. Dryer's revision of the six-way typology into SV/VS and VO/OV (Dryer 1997, Dryer 2013c) and his definition of basic order in terms of frequency (Dryer 2013a) in particular provide a simple and empirical conceptual apparatus for further work not only on inter-language, but also intra-language variation.

2.3 The generative approach

2.3.1 *Aspects* to Minimalism

In the foundational work of transformational (or generative) grammar, 1965 *Aspects of the Theory of Syntax*, Chomsky (1965: 16) postulates that every sentence has a "deep structure" (the semantic component) and a "surface structure" (the actual phonetic realization). These may be identical, but are generally not, and in such case, the latter is the result of a set of operations called "grammatical transformations" applied to the former (Chomsky 1965: 16-17). This distinction – which Chomsky describes as "the

central idea of transformational grammar" (Chomsky 1965: 16) – dominates the analysis of constituent order in most branches and offshoots of generative grammar. In specific terms, this translates to the assumption that sentences are created using two sets of rules, phrase structure rules (such as the classic example of S → NP VP) which generate the base form of the sentence and transformation rules which produces various types of syntactic structures. In the terms of variation in constituent order, this translates to the hypothesis that each language has a base-generated deep constituent order and any variation in the surface constituent order, whether within a language or among languages, is the result of transformations, typically movements. The Wh-movement is perhaps the most notorious example of constituent order variation within a language that is explained by transformations (see Chomsky 1964: 37-38 for one of its early formulations); passive transformations, Subject Raising and Clefting are but a few others. McCawley's analysis of English deep constituent order as VSO as opposed to its standard surface configuration SVO (McCawley 1970) is an example of the extremes to which such analyses could be taken.

The $\bar{X}$ (X-bar) theory (Jackendoff 1977) is a 1970s development of the generative theory of phrase structure. It introduces the concept of "heads" and "functional projections" and assigns a crucial role in sentence production to lexical categories: nouns (N), verbs (V), adjectives (A), Determiners (D), prepositions (P) etc. – all generically referred to as X – serve as heads of phrases. X combines with 0 or more complements or adjuncts to create an $\bar{X}$ or X' (intermediate projection); X' combines with a specifier to form a X".

In terms of analysis of constituent order variation, the $\bar{X}$ theory continues in the tradition of earlier works; as such, it does not allow the generation of sentences with VSO constituent order typical for languages such as Irish or Classical Arabic, since the fundamentals of the theory assume that nothing can come between the verb and its object, i.e. break up the Verbal Phrase (Carnie 2013: 300). The $\bar{X}$ theory therefore postulates that the deep structure (or D-structure in $\bar{X}$ theory terminology) of such sentences is SVO and the VSO order comes about as a result of verb movement (Carnie 2013: 301-303).

The Principles and Parameters (P&P) model, a 1980s refinement of generative grammar, abandons the concept of transformation operations, save for "a single rule Move-α that constitutes the transformational component", Chomsky 1981: 18)[2] where α stands for any lexical category; movements like Subject Raising are now merely instances of Move NP. Even phrase structure rules (Chomsky 1981: 16 footnote 19, Chomsky 1995: 25) are out the window; instead, P&P assumes the existence of Universal Grammar which "consists of interacting subsystems … of principles" (Chomsky 1981: 5, Chomsky 1995: 3-4) and a set of binary parameters (including phrase structure rules, cf. Chomsky 1995: 25). These parameters determine the actual properties of a language

2 14 years later, however, Chomsky asserts that "The transformational rules still exists, but only as principles of UG, freely applicable to arbitrary expressions" (Chomsky 1995: 25).

(Chomsky 1981:7) and consequently, the grammar of any language "can be regarded as simply a specification of values of parameters of UG" (Chomsky 1981: 31). The purpose of P&P is then to "identify and clarify parameters of UG" (Chomsky 1981: 6). Some P&P analyses propose the existence of a directionality parameter which determines "the relative order of a head and its complement" (Mahajan 2003: 217, see also Chomsky 1995: 53), others offer a more complex picture: for example, one work seeks to identify the parameters behind the VSO order in Welsh (Roberts 2005) and finds that "AgrS (subject agreement) has a weak D-feature and a strong V-feature. These are the parameter settings that give rise to VSO order" (Roberts 2005: 43). Additionally, P&P extends the concept of S-structure with the notion of interfaces: S-structure now consists of Phonetic Form (PF) which specifies the sound realization of a sentence and Logical Form (LF) which specifies its meaning which are both said interfaces in that they "have an interpretation in terms of the sensorimotor systems" (Chomsky 1995: 21).

The Minimalist Program (MP, Chomsky 1995) is an extension of P&P: where P&P's aim is description and comparison (cf. Chomsky 1995: 6), MP seeks to study the "computational system for human language" (Freiding and Lasnik 2011: 1). Much of the work on minimalism is therefore occupied with the question of what exactly is the nature of these operations and what others there are (cf. Hornstein 2009) in order to determine "To what extent is human language a 'perfect' system?" and "To what extent is the computational system for human language optimal?" (Freiding and Lasnik 2011: 2). As such, Minimalism is not concerned with language per se, but rather with that part of the human mind/brain that produces language and how it does it (Hornstein 2009: 178-180). Nevertheless, P&P remains the dominant paradigm: in other words, Minimalism is what is being done, P&P is how to do it (Boeckx 2006: 16), except with some changes. For one, in Minimalist P&P, "inflected words are not created in syntax but introduced pre- or postsyntactically in fully inflected form" (Zwart 2017: 33). More importantly, however, the $\bar{X}$ theory of sentence formation is now replaced with Bare Phrase Structure (Chomsky 1995: 249, 335), the primary components of which are the operations Merge and Move (an, on occasion, Agree).

Contemporary generative – or, as its proponents insist on referring to it, theoretical – linguistics is thus a free-for-all of competing or complementing theories (e.g. incorporation theory, cf. Nikanne 2017) and research projects, some attempting to square the Minimalist circle, some adhering to the stricter formulations of P&P. One of these projects is particularly relevant to this work, as it devotes special attention to the analysis of constituent order and its variation.

2.3.2 The Cartographic Project

The Cartographic Project is a development of the $\bar{X}$ theory devoted to the study of syntactic structures, more specifically, "the discovery and mapping out of the functional structure of natural language sentences" (Cinque 2002b: 3). This functional structure

combines with the lexical structure to create clauses and phrases (Cinque and Rizzi 2010: 52) and is in itself "a very complex and pre-existing object, consisting of a very large number of heads, ordered among each other and specialized in their function" (Benincà and Munaro 2010b: 4). At the very least (Cinque and Rizzi 2010: 52), the functional structure of sentences consists of

I. Verbal Phrase (VP),
II. Inflectional Phrase (IP) which contains heads "corresponding to concrete or abstract morphological specifications of the verb" and licenses case agreement and other features (Rizzi 1997: 281), and
III. Complementizer Phrase (CP) which "hosts topics and various operator-like elements such as interrogative and relative pronouns, focalized elements etc." (Rizzi 1997: 281)

with further subdivisions "into more articulated hierarchical sequences of functional projections" (Cinque and Rizzi 2010: 52).

The Cartographic Project is devoted to the analysis of the "fine structure" (Rizzi 1997) of these sequences and a substantial portion of that analysis focuses on a particular part of the CP, the so-called left periphery. The analyses of the left periphery are especially relevant to the study of constituent order and information structure, as the left periphery is where many European languages put their topicalized and focalized constituent, whether by default or as a result of a movement. Much of the work within the Cartographic Project thus goes into determining the exact structure of the left periphery in individual languages, cataloguing in detail the cross-linguistic variation of said structure and determining which of its components are universal. This is often done in terms of establishing a hierarchy of individual functional projections (questions and relativizers, but also topic and focus) and describing the positions – or "fields" – in which these can occur. To give an example, the findings of one such analysis of topic, focus and constituent order in Medieval Romance (Benincà 2006: 61) can be summarized as follows (visualization expanded with full names of heads, except for C° which stands for any CP head):

> [Force C°][Relative*wh* C°]/{$_{\text{Frame}}$[SceneSetting][HangingTopic] C°}{$_{\text{TOPIC}}$[LeftDislocation] [ListInterpretation] C°} {$_{\text{Focus}}$[I Focus][II Focus]/[Interrogative*wh*] C°}[Finite C°]

This typology of functional categories in the left periphery can also be extended to the right periphery of the sentence: while the left periphery is where topicalized and focused constituents appear in languages like English, French (De Cat 2010) and Italian (Rizzi 1997), in other languages, similar phenomena can also be found on the right (Benincà and Poletto 2004: 68). Consequently, there have appeared studies analyzing the right periphery in Catalan (Villalba 2000, Villalba 2011), Bulgarian (Krapova and Cinque 2008) and even Maltese (Čéplö 2014) using the concepts of the Cartographic

Project like Hanging Topic and Left Dislocation and their right periphery equivalents Afterthought (or Anti-Topic) and Right Dislocation, respectively.

In addition to the analysis of information structure, the Cartographic Project also devotes a lot of attention to the typology of constituent order in general, primarily through the study of the so-called V2 languages (Holmberg 2015), a phenomenon where "the finite verb is obligatorily the second constituent, either specifically in main clauses or in all finite clauses" (Holmberg 2015: 342). Such languages have received a disproportionate amount of attention in the generative tradition in general thanks to V2 being a prominent example of head movement (Cognola 2013: 20). This continues to be the case in many works within the Cartographic Project which examine the languages which exhibit the V2 constituent order, the properties of those languages and their version of V2 and the particulars of this phenomenon.

While it is hard to judge, especially in a wide and productive field such as modern generative linguistics, recent works like Bailey and Sheehan (2017) suggest that the Cartographic Project, especially as represented in the works of Cinque, Rizzi and Benincà, serves as the dominant paradigm in the study of constituent order for most generativists. And despite general misgivings one might have vis-à-vis the underlying theory, it is undeniable that it has produced a number of worthwhile descriptive works and even provided some theory-independent and empirically based conceptual apparatus for the study of constituent order and its pragmatic variation.

2.3.3 Constituent order and pragmatics in generativist thought

Despite his focus on structural description formulated as transformation rules, Chomsky 1965 recognizes the importance of pragmatics (or, in his words, "stylistic factors" Chomsky 1965: 11) for the variation of constituent order, noting that "grammatical transformations do not seem to be an appropriate device for expressing the full range of possibilities for stylistic inversion" (Chomsky 1965: 126). Chomsky resolves this conundrum by claiming that the rules of pragmatically determined variation in constituent order "are not so much rules of grammar as rules of performance" and while interesting, they have "no apparent bearing, for the moment, on the theory of grammatical structure" (Chomsky 1965: 127).

The moment in question did not last long and soon generativist works began to appear dealing with "the annoying problem that languages differ from one another" (Carnie 2013: 27) in the ordering of the constituents. John R. Ross' 1967 PhD dissertation devotes some attention to the problem of free word order in Latin and other languages in the context of node deletion or tree pruning, i.e. reducing the complexity of sentences generated by existing theories of generative grammar (Ross 1967: 41). In the analysis of the various possible configurations of constituents and even components of noun phrases in Latin (the same phenomenon that captured Tesnière's attention, see *Elements*, Chapter 7, §8), Ross proposes the Scrambling Rule (Ross 1967: 75) which,

along with the Relative Clause Reduction Rule (Ross 1967: 28) permits the seemingly unlimited surface variation of words in Latin sentences. Ross admits to being unable to provide the full explanation of the phenomenon ("the problems involved in specifying exactly the correct subset of the strings which will be generated by [the Scrambling Rule] are far too complicated for me to even mention them here", Ross 1967: 77), but his observations would have profound influence on the generativist approach to constituent order variation (which is what scrambling describes): first, the term scrambling has now become a firm part of generativist terminology (cf. Corver and van Riemsdijk 1994a). Second, like Chomsky, Ross recognizes that rules like the Scrambling Rule are quite different from transformational rules; unlike Chomsky, Ross spells out why that is: "[the Scrambling Rule] can apply an indefinite number of times to its own output, every sentence will have an infinite number of derivations ... the number of trees that will be assigned to any sentence, although it will be bounded, will be very large, and there will be no correlation between the number of derived trees and perceived ambiguities, as there is in happier circumstances" (Ross 1967: 77). Ross therefore proposes that rules like the Scrambling Rule are not included in base generation or transformations, but rather are part of "a stylistic component ... [which states] language-particular output conditions ... which capture the notion of preferred order" (Ross 1967: 73, underlined in the original). And finally, Ross also recognizes that there is a degree to which individual languages allow scrambling: in other words, there is an inter-language variation in constituent order variation and free word order is but one end of the scale. He therefore proposes that there exists a universal language-independent skeleton rule which turns scrambling on or off and language-specific rules then determine to what extent scrambling can be applied (Ross 1967: 78-79).

The idea of a skeleton rule allowing variation in constituent order took hold in generativist thought in the form of division of languages into one of two categories – configurational and non-configurational (Chomsky 1981: 127-135, Hale 1983). Chomsky's original description framed in terms of P&P is as follows (Chomsky 1981: 132):

> The obvious analogue of the rule Move-α for Japanese is the rule (6):
> (6) Assume a GF. (Chomsky 1981: 129; GF = Grammatical Function)
> ...
> Summarizing, Japanese is non-configurational, English configurational. Thus, GFs are not represented in D- and S-structures in Japanese in terms of the formal structures, but are assigned randomly to D-structures and by (6) to S-structures.

This formulation summarizes one of the two generativist approaches to scrambling that had developed since Ross' day, the base-generation approach. It argues that variation in constituent order is a syntactic phenomenon, i.e. it is generated randomly at the D-structure level (Corver and van Riemsdijk 1994b: 1). The distinction made here is between configurational languages which do not allow this random generation of constituents and non-configurational languages (also termed "flat languages" by Hale 1983: 10, since they do not have a unitary Verbal Phrase, cf. Chomsky 1981: 28) which

do; in terms of P&P, there exists a Configurationality Parameter (Hale 1983: 25-26). In contrast, the movement approach (Corver and van Riemsdijk 1994b: 2) explains variation in constituent order by different types of movements, whereby some of the literature on the movement side of the argument narrows the definition of scrambling to specific types of movements such as object shift (Broekhuis 2008 for Germanic languages, Gallego 2013 for Romance) or VP fronting (Zubizarreta 1998). Both approaches have produced much literature (see Corver and van Riemsdijk 1994 and Karimi 2003 for overviews), but so far, without any consensus in sight.

2.3.4 Discourse configurationality

While the generativist discussion of scrambling seems to be dominated by the base-generation and movement approaches, there is still a third school of thought harkening back to Chomsky 1965 and Ross 1967 which considers constituent order variation from the point of view of pragmatics. This school, best represented by Kiss (1995a), has surveyed a number of language very different from Standard Average European (Kiss 1995b: 4) and observed that "the structural role that the grammatical subject plays in the English sentence may be fulfilled by a constituent not restricted with respect to grammatical function or case in other languages" (Kiss 1995b: 3). In simple terms, this school of thought argues that languages fall into two groups: subject-prominent languages where the surface constituent order is Subject – Verbal Phrase and topic-prominent languages, where the place of the Subject can be taken by an arbitrary element bearing a particular discourse (or pragmatic) function (Kiss 1995b: 4). These languages are termed discourse-configurational and their fundamental properties are as follows (Kiss 1995b: 6):

> A. The (discourse-)semantic function 'topic,' serving to foreground a specific individual that something will be predicated about (not necessarily identical with the grammatical subject), is expressed through a particular structural relation (in other words, it is associated with a particular structural position).
> B. The (discourse-)semantic function 'focus,' expressing identification, is realized through a particular structural relation (that is, by movement into a particular structural position).

Kiss goes on to argue that while sometimes these two properties go hand in hand, they are not interdependent and so some discourse-configurational languages can display only type A characteristics, whereas others only show the type B properties (Kiss 1995b: 6). It should be noted, however, that while the fundamentals of this subset of generativist theory are framed in terms of pragmatic function, much of the explanation offered by its proponents still depends on movements (Choe 1995), such as the Focus Movement (focalization) – which is very much akin to Wh-movement in its properties – and Topic Movement (topicalization). And as with literature on scrambling, there

seems to be no consensus in generativist literature on the general properties and nature of discourse configurationality.

2.3.5 Evaluation

One's opinion of the contribution of generative grammar to the study of constituent order depends on whether one accepts the fundamental assumptions of the generativist approach regarding Universal Grammar and the dichotomy between the deep structure and the surface structure. As I noted in chapter 1, I don't, and consequently, the generative theorizing is irrelevant for the purposes of this thesis. This is not to say that there isn't anything of value in the generativist literature at all, quite the contrary: as Haspelmath (2009: 345) observes, "[m]any papers in the generative tradition first provide a fairly framework-free description of the relevant phenomena ('the data') and then go on to provide a second, framework-bound description ('the analysis')". Fabri 1993 is a prime example of this: while the framework-bound part of this analysis of agreement and related phenomena in Maltese hasn't aged well, Fabri's description of the workings of the phenomena in question is still unsurpassed in its comprehensiveness. The same applies to many works in the Cartographic Project or its offshoots (e.g. Bentley et al. 2015) and I will therefore happily refer to these and any other descriptive works of any generativist without any prejudice, but refrain from considering or even commenting on the theory.

What then is the purpose of this section, you ask? It is two-fold, I answer: first, as generative grammar and its various descendants constitute one of the dominant approaches to modern linguistics, much of its terminology has made its way even into non-generativist literature. Concepts like configurationality and head-finality/head-initiality are used in works that do not explicitly subscribe to the tenets of any of the generativist frameworks, including descriptions of Maltese (Borg and Fabri 2016), and the same is true of concepts like topicalization and focus fronting (Borg and Azzopardi-Alexander 2009) or even the entire nomenclature of a particular program within the generative tradition (as the Cartographic Project in Čéplö 2014). It is therefore important to put them in proper context in order to evaluate their utility and appropriateness. And secondly, the distinction between syntactically-determined and pragmatically determined variation in constituent order popularized, if not invented, by Chomsky (Chomsky 1965: 126) continues to be a constant theme in studies of constituent order.

2.4 Pragmatic approaches to constituent order

2.4.1 Functional Sentence Perspective (FSP)

Unlike most approaches discussed so far, the Functional Sentence Perspective (in Czech "aktuální členění větné", henceforth FSP) revolves around the idea that constituent order and pragmatics are intrinsincally linked. FSP is built on Vilém Mathesius' fundamental insight that "[t]he functional analysis of a sentence must be juxtaposed to its formal analysis" ("Aktuální členění věty je třeba klásti proti jejímu členění formálnímu." Mathesius 1939: 171; see Firbas 1992: 22 for the English terminological choice). Expanding on previous work by Weil (1844) and von der Gabelentz on the distinction between grammatical subject and "psychological subject" ("das psychologische Subjekt", von der Gabelentz 1869: 378), Mathesius establishes a two-way division of sentence in terms of its communicative effect: the "theme", defined as "a thing about which we assert something" ("to, o čem něco tvrdíme", Mathesius 1961: 91) and "what we say about the theme is the nucleus or the enunciation" ("to, co o základu tvrdíme, je jádro výpovědi neboli vlastní výpověď", Mathesius 1961: 92).[3] This division, for which Mathesius' successors (Firbas 1957) established the terms "theme" and "rheme", is the cornerstone of FSP and the fundamental principle of its theory of communication (cf. Krifka 2007: 13).

Mathesius's work was expanded on by Daneš (1959) and by Firbas (1957, 1964, 1992) who became the major theoretician of FSP. Firbas elaborated on the theme/rheme distinction by introducing two new concepts: the first is "communicative dynamism (CD), a phenomenon constantly displayed by linguistic elements in the act of communication" (Firbas 1992: 7). CD is gradual, rather than binary, and as such, it should be understood to be "the relative extent to which a linguistic element contributes towards the further development of the communication" (Firbas 1992: 8). CD is not necessarily linear, nor does it only involve syntax – in fact, Firbas 1992, like Daneš 1957, spends much time on the role of intonation and prosody in conveying degrees of CD. CD is expressed (or, in Firbas' terminology, carried) by linguistic elements which can be syntactic, morphological and even semantic (Firbas 1992: 17) and is distributed across sentences and their constituent parts, be they clauses or phrases. These are termed "distributional fields" (Firbas 1992: 15) and the distribution of degrees of CD over a sentence (and its constituent distributional fields) then determines the functional structure (or the "functional sentence perspective" proper, cf. Firbas 1992: 21) of a sentence.

In Firbas's understanding of FSP, Mathesius's distinction between "theme" and "rheme" applies to all distributional fields (thus, for example, accounting for the variation in noun-adjective order in noun phrases). Firbas expands Mathesius's analysis by

3 In English translations of Mathesius's work, "theme" is also referred to as "basis (of the statement)", see footnote 77 in Mathesius 1961: 91.

acknowledging the role context plays in the structuring of CD ("Sentences are usually embedded in the flow through context-dependent elements", Firbas 1992: 68) and uses the very same to provide a more firmly grounded definition "theme", in which the term refers to the aforementioned "context-dependent elements". For the other side of the opposition, Firbas uses the term "non-theme" (Firbas 1992: 71) and introduces its subdivision into at minimum "transitional elements", or simply "transition", and "rheme" (Firbas 1992: 71-72). In terms of CD, each of these three major components (theme-transition-rheme or Th-Tr-Rh) carries a higher degree of CD than the previous one. As Firbas makes clear, only rheme proper and transition proper need to be implemented within each distributional field (Firbas 1992: 72).

In terms of analysis of constituent (and word) order and its variation, Firbas starts out with Mathesius's formulation of "word-order principles" which include "the principle of grammatical function, the principle of coherence of members, the principle of FSP, the principle of emphasis and the principle of sentence rhythm" (Firbas 1992: 117, see the original formulation in Czech in Mathesius 1961: 180-191). Firbas combines the first two into one and then renames the third into "FSP linearity principle", as in this context, it reflects the ordering of sentence constituents according to the Th-Tr-Rh order (Firbas 1992: 118). The principle of emphasis (or "the emotive principle" in Firbas' terminology) comes into play in contexts where a contrastive or "marked" reading is desired (Firbas 1992: 120-121) and it invariably arranges the words in question in the order contrary to that deviates "from syntactic patterns regarded as normal" (Firbas 1992: 122). And finally, sentence rhythm "produces a certain pattern of heavy and light elements" and typically combines with the emotive principle to produce a marked order (Firbas 1992: 119). These four principles constitute the fundametal theory of constituent order variation in FSP and its offshoots, with FSP as the primary factor in languages like Czech (Panevová et al. 2014: 209-210).

2.4.2 Functional Generative Description (FGP)

Functional Generative Description (henceforth FGP after the original Czech term "funkční generativní popis") was originally an attempt to reconcile the generative theory of sentence production with FSP (Sgall 1967a). Building on the fundamental distinction between semantics (contents of the mind as encoded in the lexicon, cf. Sgall 1967b) and the acoustic or written form of an utterance, FGP is devoted to the description of the process by which the former is turned into the latter (Sgall, Bémová et al. 1986: 114). That description is performed on five different levels, each of which has its own fundamental units and rules of combining them (Sgall, Bémová et al. 1986: 114). The levels of analysis in FBP are conceived as sets of sentence notations (Sgall, Bémová et al. 1986: 111-112) and they are (ordered by decreasing depth or, in Sgall's terminology, height) as follows:

1. Tectogrammatic level (semantics),
2. surface syntax,
3. morphemes,
4. (morpho-)phonology, and
5. phonetics.

The first (highest) two are the primary focus of FGP research and while they are notionally equivalent to Chomsky's deep and surface syntax, this is as far as the generative part of FGP goes. The original formulation of FGP provided a description of the actual generative component of sentence production and the role of FSP in it (Sgall 1967a: 214-220, Sgall et. 1980: 90-110) in terms of a phrase structure grammar (and related algebraic operations), but this was ultimately abandoned in favor of dependency analysis (Sgall 1967b), citing, inter alia, difficulties of describing the structure of languages with free word order using phrase structure grammars ("Frázová gramatika ostatně naráží na potíže u jazyků s tzv. volným slovosledem", Sgall 1967b: 362). Consequently, the description of both the tectogrammatic level and the surface syntax of a sentence is framed in terms of dependency, i.e. governor-dependent relationships (Sgall et al. 1980: 16) as understood by Tesnière.

FGP combines FSP and Tesnière's ideas on verb centrality and postulates the verb as the delimiter of thema and rhema (or topic and focus, Sgall et al. 1980: 12). At the same time, it argues against communicative dynamism as the primary criterion in determining constituent order and makes a strong case for a language-dependent default order of certain types of verbal dependents ("systémové uspořádání", Sgall et al. 1980: 17, 77) and "contextual anchoring" ("kontextové zapojení", Sgall et al. 1980: 17) as alternative explanations for the observed variation. The latter also serves to provide updated definitions for the concepts of "topic" and "focus" and analysis of information structure.

In its most prominent contemporary role, FGP provides the theoretical framework for the Prague Dependency Treebank (PDT, Bejček et al. 2013) and its associated projects, like the valency dictionary of Czech verbs (VALLEX, Lopatková et al. 2017) and a description of Czech syntax (Panevová et al. 2014) based on PDT.

2.4.3 Evaluation

Contemporary FSP as a subfield tends to focus more on information structure, including its interaction with discourse and units of meaning longer than a sentence (Dušková 2015, Vaculíková and Jurka et al. 2015), yet along with FGP, it continues to be the dominant paradigm in the study of constituent order in Czech linguistic tradition. This is especially true if the object of the study is modern day Czech (Panevová et al. 2014), its diachronic varieties (Zikánová 2009) or other Slavic languages (Krejčová 2016). Recent years have seen FSP analyses of a number of languages other than Czech ranging

from Italian to Indonesian (Vaculíková and Jurka et al. 2015), chief among them the extension of the principles of multilevel annotation of PDT to Arabic in the Prague Arabic Dependency Treebank (PADT, Hajič et al. 2004). The former analyses also offered refinements of the underlying theory, like the concept of "hypertheme" and the idea of rhematization of themes (cf. Vaculíková and Jurka et al. 2015: 49). Outside of Czech scholarly milieu and its Slavic environs, FSP/FGP as a framework has attracted little attention (Panhuis 1982, Skënderi 1997), with the possible exception of references to its fundamental ideas and their elaboration in works on functional linguistics (Halliday 2014, Givón 2001a and 2001b).

There is, however, one area where where FSP has left an indelible mark on modern linguistics, the study of information structure: foundational works by Mathesius (1961 in its English translation) and Firbas (1964) are credited with establishing the subfield (Féry and Ishihara 2016b: 3). Its basic terminology, redressed and redefined multiple times (e.g. topic-comment or topic-focus) and its fundamental ideas like context-boundness (Krifka and Musan 2012) have become a firm part of modern linguistic terminology (Féry and Ishihara 2016a).

2.4.4 Categorical and thetic judgments

Mathesius elaboration on Weil's "march of ideas" (Weil 1844: 23) and von Gabelentz's theory of "psychological subject" (1869: 378) may be the most prominent extension of philosophical ideas on the nature of subject and predicate to modern linguistics proper, but it is far from the only one, nor is it the first one. In fact, long before Mathesius argued for the separation of grammatical and functional analysis of the structure of sentences, the Austrian philosopher Anton Marty devoted much attention to the relationship between grammar, logic and psychology in judgments (Eisenmeier et al. 1918). Like Mathesius, Marty analyzed the contrast between what he termed "logical subject" and grammatical subject, i.e. the contrast between semantic roles and their syntactic realization. Unlike Mathesius, however, Marty attempted a more general analysis by also considering subjectless (and predicateless; cf. Marty 1895: 298) sentences and other similar syntactic phenomena and attempting their classification.

Marty's investigation elaborated on the previous work by his teacher Franz Bertano and culminated in the insight that there exist two type of sentences: those that do not express judgments (in the philosophical sense), like interrogative or imperative sentences (Marty 1897: 189) and those that do. The latter group is then further divided into two fundamentally different types : first, there are sentences which consist of a subject and a predicate (whether a copula or not; cf. "Aussageformel mit Subjekt, Prädikat und Kopula (oder dem Äquivalent derselben)"; Marty 1897: 180). These actually contain two judgments: one about the existence of the subject, the other about a property of the subject (Marty 1897: 178). Marty refers to these as "compound judgments" ("Dopperurteil"; Marty 1897: 180) or categorical judgments.

Along with those, however, there is a type of sentences which only contain one single judgment ("einfache Urteile", lit. "simple judgments"). Referring to previous work on judgment structure and deixis, Marty argues that in constructions like those in German that feature *Es gibt* ("there is"), this phrase "means nothing in and of itself" ("bedeutet für sich allein gar nichts"; Marty 1894: 847) and that like prepositions or case suffixes, it is "synkategorematisch" or, in modern terminology, synsemantic. The same then applies to German *es* "it" in sentences like *es regnet* "it's raining" or their equivalents in French like *il y a*, lit. "he there has", where *es* as well as *il* and *y* have all lost their original deictic meaning (Marty 1894: 847-848, Marty 1895: 296). As such, they do not contain actual subjects and the only judgment expressed is the predicate, hence their classification as simple judgments. To those, Marty also adds one more type of sentences, sentences which "only appear to have a subject and a predicate" ("nur scheinbar Subjekt und Prädikat haben"; Marty 1895: 298) but whose "actual meaning is "a simple recognition or negation" ("ihre eigentliche Bedeutung ist eine einfache Anerkennung oder Verwerfung"; Marty 1895: 298). These are sentences like "All triangles have three sides" and "No sound is a color". Their semantics makes them equivalent to other types of single judgments, their syntax is that of categorical judgments; hence their classification as pseudocategorical ("pseudokategorische Sätze") or categoroid sentences ("kategoroide Sätze"). They, along with existential and impersonal sentences (Marty 1895: 301-302), are subsumed under the label of "thetic utterances" ("thetische Aussagen"; Marty 1895: 298), contrasted to that of categorial judgments.

Marty's insights remained largely ignored in modern linguistics until Kuroda (1972) who explicitly connected Marty's classification of sentences with the study of information structure tying them to the concepts of "topic/thema" and "comment/rhema" (noting the role of the Prague linguistic circle in developing these terms, Kuroda 1972: 157-158) and "focus" and "presupposition" (Kuroda 1972: 159). Kuroda applies Marty's theory to the study of information structure in Japanese (more specifically, focus, cf. Nakagawa 2018: 30-31), using the distinction to account for the behavior of two seemingly synonymous particles. His larger contribution lies in the application of Marty's categorical-thetic distinction to the study of information structure and establishing that only categorical sentences can feature topics and thus the topic-comment (theme-rheme, topic-focus etc.) distinction. Thetic sentences, by their very nature, provide new information (information non-derivable from the context) only and thus cannot be divided into topic/comment or thema/rhema. As such, they are not sensu stricto subject to information structure and if they are considered to be, they are invariably equivalent to sentence or predicate focus (Nakagawa 2018: 31, Zimmermann 2016: 316-317).

Since Kuroda, the literature on thetic sentences has expanded somewhat, though, it would seem, not in step with literature on other aspects information structure. Sasse 1987, Sasse 1995 and Rosengren 1997 have elaborated on the theory, works like Güldemann 2010 applied the analysis to the Tuu language family of South Africa. The last

named work in particular illustrates the way in which the concept of theticity has been incorporated into the study of information structure in terms of focus.

Sasse (1987) offers a detailed refinement of the theory: first, he posits further classification of thetic expressions into "entity-central" and "event-central" (Sasse 1987: 527) where existential (or "presentational") expressions fall under the former, and impersonal sentences describing weather (see Marty 1894: 847-848) classify as the latter. Secondly and more importantly, Sasse provides the following typology of thetic sentences (Sasse 1987: 566-567):

1. EXISTENTIAL STATEMENTS (in a wider sense; presence, appearance, continuation, etc., positively and negatively)
2. EXPLANATIONS (with or without preceding questions such as 'what happened?', 'why did it happen?', etc.)
3. SURPRISING OR UNEXPECTED EVENTS
4. GENERAL STATEMENTS (aphorisms, etc.)
5. BACKGROUND DESCRIPTIONS (local, temporal, etc., setting)
6. WEATHER EXPRESSIONS
7. STATEMENTS RELATING TO BODY PARTS

As Sasse (1995: 4) notes, the concept of theticity "has been addressed under various aliases". For example, theticity is precisely what Petrova and Solf (2008) describe in their observations regarding constituent order in the Old High German translation of Tatian. They note that verb-initial sentences indicate shifts in the narrative (cf. Sasse's "episode-opening function", Sasse 1995: 16) and this order often occurs in orally transmitted literature, such jokes (Petrova and Solf 2008: 333-335); as such, verb-first sentences highly correlate with verbs of motion (Petrova and Solf 2008: 334). Cichosz (2010: 87) further observes that the same phenomenon occurs in Polish, Yiddish and Finnish and proposes that this may reflect its universal nature. Sgall (Sgall et al. 1980: 39-44) discusses the concept without naming it (but with reference to Kuroda 1972) in the context of verbs "introducing to the scene" or "stating existence" ("uvádění na scénu ... konstatování existence", Sgall et al. 1980: 40), but rejects their special nature, thus incorporating them into his understanding of functional sentence perspective. And finally, Manfredi (2017) uses the concept of theticity in the study of pain expressions in Arabic varieties, including Maltese.

2.5 The quantitative approach

The name of this final school of thought may give a careful reader pause, since quantitative analyses of constituent order are the fundament upon which typological approaches to the study of the same rest – Greenberg (1966), Hawkins (1983) and Dryer (1992) all count numbers and cite percentages. Theirs are, however, inter-language

studies, ones where the basic order was established for a number of languages and those are being compared. In contrast, quantitative studies referred to in this section are studies of constituent order within a single language, whether aiming to determine what is the basic one or analyzing the variation.

Despite the title I gave this section, these cannot actually be considered a unified approach per se, though there have been attempts to unite them into one (Köhler 2012); this category is rather a wide umbrella of more or less ad-hoc investigations. And while such studies have increased in number in the last two decades in step with the increased availability of large-scale corpora, they are far from a new phenomenon: for example, as early as 1967, Ludmila Uhlířová quantitatively analyzed the distribution of possible configurations of subject, object and predicate in Czech on an ad-hoc corpus of 5400 sentences (Uhlířová 1967), finding that (S)VO, O(S)V and VOS are the most frequent ones, with VSO and OSV rare, the latter extremely so. With the renewed interest in corpus linguistics and empirical methods in linguistics in general, quantitative studies began to appear in larger numbers, examining a number of languages like Italian (e.g. Sornicola 1994), English (Arnold et al. 2000, Wasow and Arnold 2003), French (Thoullier et al. 2014), Czech (Rysová and Mírovský 2014), German (Heylen 2005, Bader and Häussler 2010) and even Guaraní (Tonhauser and Colijn 2010). Quantitative studies of intralanguage variation have also become increasingly popular in diachronic linguistics, where one work examines the order of subject and predicate in Old Czech (Zikánová 2009), others analyze the position of object in Old English (Taylor and Pintzuk 2012) and the constituent order and information structure in Old Spanish (Sitaridou 2011).

The use of corpora and treebanks for (mostly diachronic) research has been on the rise even in works based on frameworks which typically shun empiricism in any of its forms: one recent generativist work uses historical corpora to study a number of phenomena related to changes in constituent order in the history of Spanish (Poole 2017), another examines the constituent order in Danish subordinate clauses based on a corpus of spoken Danish (Jensen and Christensen 2015) and another provides a detailed quantitative analysis of the development of Latin constituent order using an extensive corpus of Latin texts (Danckaert 2017). With the increasing availability of corpora and treebanks and the refinement of methodology, the subfield of quantitative constituent order analysis can only be expected to grow in the future.

2.6 Summary

In this chapter, I have shown the importance of constituent order as an object of study in linguistics, both for general descriptive purposes on the level of individual languages, as well as within the larger task of describing and account for the differences between languages (language typology). The former presupposes the latter and consequently, a detailed description of constituent order should be provided for each and every one

for the world's languages. In this thesis, I endeavor to do just that for Maltese and as a preliminary, in the next chapter I will examine how the approaches to the study of constituent order have been used for that purpose in previous works on Maltese.

3 Maltese constituent order: state of the question

3.1 Introduction

3.1.1 General

For a numerically small and geographically and culturally marginal language, Maltese boasts a long and rich tradition of scholarly interest. This is evidenced, inter alia, by the fact that the first grammatical description of Maltese worthy of the name, De Soldanis' 1750 *Nuova scuola di grammatica per agevolmente apprendere la lingua punica – maltese* (published in de Soldanis 1750), predates the first actual printed book in Maltese (Francesco Wzzino's translation of the Catholic Catechism titled *Taġhlim Nisrani* published in Rome) by two years. The next century and half sees the publication of a number of grammars and grammatical treatises by both native scholars (Vassalli 1791, Vassalli 1827, Vella 1831 and Panzavecchia 1845) and foreign ones (Stumme 1904, Roudanovsky 1910 and Roudanovsky 1911). Like many early grammars, however, these focus mostly on phonology and morphology, either for their own sake or in order to study the relationship of Maltese to other languages. Those early works that do aspire to describe Maltese syntax – Vassalli 1791, Vassalli 1827, Vella 1831 and Panzavecchia 1845 – discuss it briefly and mostly focus on issues such as agreement (Vassalli 1791: 213-215, Panzavecchia 1845: II.2-4), verbal auxiliaries (Panzavecchia: II.9-14) and verbal valency (Vassalli 1827: 140-146, Panzavecchia 1845: II.6-9). As such, they do not – with one notable exception, that of Vella 1831 – address the issue of constituent order; the closest they get to it is examining word order within the noun phrase (Vassalli 1827: 128 or Panzavecchia 1845: II.15-16).

In this chapter, I look at those early and modern descriptions of Maltese that offer an opinion on the constituent order of Maltese, adding a discussion of the details and evaluating the merits of that opinion.

3.2 Vella 1831

3.2.1 Overview

Francis Vella's 1831 *Maltese Grammar for the Use of the English* is the first grammar of Maltese (and the only one written before the 20th century) that describes the constituent order of Maltese. As it was written for the purpose of "the improvement of our language [and] our desire of affording some means to such english (sic) individuals, as may be desirous af (sic) learning it" (Vella 1831: 7), it is structured as textbook with chapters and sections divided into classes and lessons. The second part of the grammar (Vella 1831: 224-297) is devoted to syntax which Vella describes as being "three-

fold, namely of *concord*, *government*, and of *construction*" (Vella 1831: 224, italics in the original). As with Vassalli's and Panzavecchia's grammars, however, many of the chapters nominally covering syntax actually focus entirely on morphological issues.

It is in this second part of Vella's grammar that we find two remarks on the nature of Maltese constituent order interspersed among the description of other phenomena. They are, by and large, the product of practical focus of the grammar and its contrastive bend: Vella regularly provides comparisons of the structure of Maltese to that of English in order to highlight where the former differs significantly from the latter and where they are similar.

The first observation on the constituent order in Maltese can be found at the very beginning of "Part the Second" in chapter I, section I, lesson I which covers the nominative case i.e. the subject (Vella 1831: 224-225). I reproduce the section in its entirety and original orthography below:

> The nominative is commonly placed before the verb.
>
> The sailor crosses the sea, and the peasant ploughs the land.
> *Il baħri jaksam il baħar, u il bidui jaħrat l'art.*
>
> But the nominative goes after the verb when the verb is preceded by li, that,
> when that is relative to a noun in the accusative case, as.
>
> The house that my father built.
> *Id dâr li bena missieri.*

In chapter VI, section I, lesson II (Vella 1831: 253), Vella also has a word or two to say about the position of the adverb as well:

> Adverbs of place go mostly after the verb in both languages; as,
>
> | | |
> |---|---|
> | I am going thither. | Sejjer hemm. |
> | He was below. | Chien isfel. |
> | She comes from thence. | Gejja min hinn. |

In addition to these observations, Vella also comments on issues of word order, noting that the adjective always follows the noun (Vella 1831: 231) as do possessive pronouns (Vella 1831: 235) and adverbs with respect to adjectives (Vella 1831: 252), thus in effect describing the Maltese noun phrase and adjectival phrase as head-initial.

3.2.2 Summary and evaluation

Translated to modern terms, Vella describes the default order of a Maltese sentence as SV in main clauses, but VS in (one type of) relative clauses, those modifying the object

(Vella's "noun in the accusative case"). The identification of those two types of variation is Vella's primary contribution to the study of constituent order in Maltese.

3.3 Sutcliffe 1936

3.3.1 Overview

Edmund Sutcliffe's 1936 *A Grammar of the Maltese Language* is not only the first grammar of Maltese to use the modern Maltese orthography (the orthography of *Tagħrif fuq il-Kitba Maltija* 1924), but it is also the first truly modern and comprehensive grammar of Maltese, in many ways still indispensable. In the introduction, Sutcliffe sets out to "present the grammar and syntax of this most interesting Semitic language" (Sutcliffe 1936: v) and with regard to the latter, does so consistently: each of the five chapters devoted to the morphology of individual word classes (chapter II through VI) concludes with a chapter on the syntax of the word class in question. In one of those, chapter IV on the adjective, he also addresses word order, noting in that "The normal position of adjectives is ... after their noun", but observing that "In less prosaic expressions there is, however, a certain latitude" and that "Superlatives quite regularly precede their noun" (Sutcliffe 1936: 63).

Sutcliffe devotes the entire chapter IX to the syntax of the sentence. This chapters covers five phenomena, two of which have to do with constituent order. First, Sutcliffe describes the "Order of sentences" as follows (Sutcliffe 1936: 210, emphasis and italics in the original):

> (a) Order of Sentences. The normal order of words in a sentence is verb, subject, object, but this order is frequently altered for reasons of euphony or emphasis. **Żagħżugħ xerred għajdut; sama' b'dil-biċċa missieru** *a youth spread a rumour; this behavior came to the ears of his father* …
> In subordinate clauses the usual order is that the verb precedes its subject. **Xi kus li bil-ħeffa tiegħu, meta daħal ilma, żamm fil-wiċċ** *some pitcher which on account of its lightness remained on the surface when the water entered*; **bħalma ħasbu xi uħud** *as some have thought* …

Second, Sutcliffe notes the large degree of variation in constituent order, but explains it in vague terms only, save perhaps for one type of structure which he discusses in the very next section (Sutcliffe 1936: 210-211, emphasis and italics in the original):

> (b) The independent nominative. The *nominativus pendens* is a common feature of Maltese style. It may be the anticipated subject of a subordinate clause: **dan kollu tafu mnejn hu ġej?** *do you know where all this comes from?* ... Or it may represent the object, direct or indirect, of the main or a subordinate verb. In this case it is resumed later in the sentence by a pronominal suffix. Thus: … **lil dawn nafuhulhom għal tliet ħwejjeġ** *to these we are grateful for three things* …

3.3.2 Summary and evaluation

Sutcliffe's observations introduce two themes that would continue to appear in the study of constituent order of Maltese: first, there are the two competing views of default constituent order in Maltese – SVO and VSO – where Sutcliffe is the first to argue for the latter. Second, his discussion of *nominativus pendens* introduces the concepts of topicalization (see the second example in b) above) and dislocation (the first example in b) above), if not by name. Sutcliffe's choice of the term is interesting in both the insight that it provides and in its obvious inadequacy: as he himself notes, the word or phrase in question can be a direct object – which, when marked with lil, he elsewhere refers to as accusative (Sutcliffe 1936: 203) – so the actual term to be used here is *casus pendens*. The distinction is important: if one wishes to stick to the terminology of the Classics, *nominativus pendens* is disconnected from the sentence syntactically, where *casus pendens* is usually interpreted as pragmatically determined variation in constituent order (Rubio 2009: 205-206).[1] While Sutcliffe does not make the distinction, he is the first one to point out the existence of the two phenomena in Maltese.

In general terms, Sutcliffe's contribution to the grammatical description of Maltese remains relevant to this day, especially in the field of morphology where he provides some of the most detailed description of the verbal system of Maltese. When it comes to syntax, however, his insights are limited.

3.4 Aquilina 1959

3.4.1 Overview

Joseph Aquilina's 1959 *The Structure of Maltese: A Study in Mixed Grammar and Vocabulary* is the first truly comprehensive study of Maltese phonology, morphology and syntax; moreover, it is the first grammar of Maltese which takes into account its hybrid structure. Part IV (Aquilina 1959: 323-351) is dedicated to "The Syntax of Semitic and Mixed Maltese" and primarily consists of a detailed list of "syntactic combinations" which form "the core of the various phrase and sentence constructions" (Aquilina 1959: 324). In actuality, Aquilina discusses two different phenomena: first, he provides a comprehensive overview of the structure of various types of phrases (or perhaps an overview of the valency of individual parts of speech, Aquilina 1959: 325-340) where he describes both the composition and the order of elements within said phrases in terms of relationships between parts of speech (e.g. nouns and articles or verbs and suffixed pronouns). Secondly, and more importantly for our purposes, Aquilina discusses

1 For a detailed analysis of the question whether *nominativus pendens* (or Hanging Topic Construction) and *casus pendens* (or Left/Right Clitic Dislocation) exist as separate phenomena in Maltese, see Čéplö 2014: 209-212.

the "rules of syntactic combinations governing the structure of the sentence" (Aquilina 1959: 341) in terms of relationships between subject (**S**), verb (**V**), direct object (**O**), indirect object (**O**), objective (**o**), suffixed pronouns (**prn.sf**) and function words (*ma*, *lil* etc.). He lists the following possible configurations:

(1) **S+V+O** [**O** is udentified] …
(2) **S+V+lil+O** [**O** is identified, singled out] …
(3) **V+S** [combination occurs (i) in emphatic or high-flown literary language; or (ii) in subordinate clauses in which, however, the usual order S+V, though less common and less idiomatic, is also heard] Ex. (i) **ˌqal ˌħu:k li** etc. 'your brother did say that' etc. (ii) **meta_daħal ˈħu:k** (also meta ˈħu:k daħal) 'when your brother entered'; but invariably **S+V** when the 2nd constituent is bound up with a following word. Ex. **meta ˌħu:k daħal_idˈda:r**, never **meta daħal ħu:k idˈda:r**. [Footnote 1: However, if **idˈda:r** occurs as a divisible constituent, or, as one might say, additionally, we can also have a subordinate clause with the combination **V+S+**. Ex. **meta daħal ˈħu:k** +(after a slight breath pause) **idˈda:r**, with the last two constituents (a+N) loosened from the main phrasal structure.]
(4) **S*+S(+/-)+V+O+(-S*)V**. [1st constituent at the beginning of the statement is S of 2nd V – It is known in Grammar as *Nominativus pendens*] Ex. **ˈdawn yien nɣeyḍ̬ ʃiˈfissru** also **ˈdawn nɣeyḍ̬ ʃiˌfissru ˈyien** [these, I say what they mean] 'I say what these mean' similarly (4) **O*+S(+/-)+V†prn.sf.*** Ex. **(Dawn) il_ˈkoṭba ˌʃtraythom** [these] the books I bought them] 'I bought these books.' Note that **O = prn.sf**.
(5) **S+V+O lil†prn.sf** [(i) variant of **S+V†prn.sf** (= No. 1 above) for emphasis, or (ii) when the 2nd constituent is followed by more than one **o prn. sf.** Ex. (i) **ˈraytek** 'I saw you' but for emphasis and singling out **rayt ˈli:lek** 'it was you (i.e. not some one else) that I saw'; …

3.4.2 Summary and evaluation

As evidenced from the list of findings in 4.4.1, Aquilina stakes a position contrary to that of Sutcliffe, arguing for SVO as the default constituent order in main clauses and for VS as a variation thereof used not for pragmatic or specific communicative purposes on sentence level, but rather as a convention employed by specific genres or types of texts. On the constituent order in subordinate clauses, as well as on the phenomenon of *nominativus pendens*, Aquilina agrees with Sutcliffe, expanding on the observations he provided in two aspects: firstly, he notes the role that the object marker *lil* plays (configurations 2 and 5 in 4.1 above), although whether he's correct on what that role is remains doubtful. Secondly and more importantly, Aquilina attempts to account for the observed constituent order variation in subordinate clauses in specific terms, this time suprasegmental phonology of the clause (configuration 3, example ii above).

3.5 Vella 1970

3.5.1 Overview

In 1970, Joseph Vella submitted to the Royal University of Malta an MA thesis supervised by Joseph Aquilina and titled *A Comparative Study in Maltese and Libyan (Benghazi Dialect): Phonetics, Morphology, Syntax and Lexicon*. The title is somewhat misleading, as this work is actually a detailed contrastive study of Maltese and Benghazi Arabic phonology (Vella 1970: I.1-44) and morphology (Vella 1970: II.45-350), accompanied by an alphabetically arranged glossary of the two languages (Vella 1970: III.1-115) and a bilingual phrasebook in six lessons (Vella 1970: IV.1-19) Further contrary to the subtitle, the thesis has little to say on the syntax of either Maltese or Benghazi Arabic; the only part that explicitly does so – chapter 8, section (s), titled "The Syntax of the Maltese and Libyan sentences" (Vella 1970: II.302-304) – only discusses coordination (which Maltese and Libyan employ, so Vella 1970: II.303, where "the European mind would prefer the use of subordinate clauses"), the tendency of "Semitic sentences" to use direct speech instead of indirect speech (Vella 1970: II.304) and clitic doubling (Vella 1970: II.304).

There are, however, references to syntax distributed across the entire work. The one most relevant for our purposes can be found at the beginning of chapter 3 titled "The Case Endings" in section b) where Vella discusses the influence of the loss of case endings on the marking of syntactic function of words. I reproduce the entire section below (Vella 1970: II.98, original emphasis and transcription maintained):

> b) To make up for the loss of case endings in Maltese and Libyan,[2] it is the position of words in the context which shows their function. As a rule, the subject of the sentence is placed close to the verb and usually follows it, as is the Semitic custom.
> Exx.
>
> | Maltese | Libyan |
> |---|---|
> | 'daħal il-'ħai'ya:t = | 'dxal (xaʃʃ) el xai'ya:ŧ |
> | the tailor entered; Lit. entered the tailor | |
> | ħab'bitu 'martu 'f'ʔalba = | ħab'beta 'mrätah fi: 'galbəha: |
> | his wife loved him in her heart; lit. loved him his wife .... | |
> | fiz-'zmiən il-ʔa'di:m | fi: z-zə'ma:n el-ge'di:m |
> | 'kiən 'emm sul'ta:n = | 'ka:n 'fi:h sul'ŧa:n |
> | a long time ago there was a king. | |
>
> Remark:- Due to foreign influence, however, the subject
> often precedes the verb. Exx.

2 In effect, Vella argues for the grammaticalization of constituent order (i.e. that subject and object are distinguished by their position), but in Neo-Arabic, not Maltese per se.

Maltese	Libyan
ir-'ra:dʒəl 'ma:r id-'da:r = the man went home.	ir-'ra:ʒəl 'ʕadda l-'ħo:ʃ
or	
'ma:r ir-'ra:dʒəl id-'da:r	'ʕadda ir-'ra:ʒəl el-'ħo:ʃ

3.5.2 Summary and evaluation

In broad strokes, we see Vella concur with Sutcliffe regarding the nature of the default constituent order in Maltese. Vella's views, however, hold little merit: while the idea that both Sutcliffe and Vella may have been unduly influenced by the other main subject of their research – Hebrew for Sutcliffe, Arabic for Vella – is easily discounted in Sutcliffe's case in light of his obvious erudition and skill, in Vella's, there is plenty of evidence, such as his naïve remarks regarding the "Semitic custom" and the "European mind". Vella's MA thesis thus remains an important source of information on Benghazi Arabic and even Maltese phonology and morphology, but sadly little of value can be gleaned from it regarding Maltese syntax.

3.6 Krier 1976

3.6.1 Overview

Syntax is conspicuously missing from the title of Fernande Krier's 1976 *Le maltais au contact de l'italien. Etude phonologique, grammatical et sémantique*; nevertheless, the slim volume does contain a chapter on the subject (Krier 1976: 63-106) which also includes a section on constituent order in Maltese. Krier's understanding of syntax is inspired by Martinet's functionalist approach (Martinet 1969 and 1970) where the basic unit of language is an utterance ("énoncé", Krier 1976: 44), defined as "a speech stream segment of variable length uttered while transmitting the particulars of an experience" ("un segment plus ou moins long de la chaîne parlée dans la transmission des données de l'expérience", Krier 1976: 44). An utterance, in turn, consists of a "linear succession of meaningful units, the monemes, thus establishing relationships which may exist between these units" ("une succession linéaire d'unités significatives, les monèmes, en déterminant les rapports qui peuvent exister entre ces unités", Krier 1976: 63).

Interestingly, Krier actually uses a proto-corpus approach, by first assembling a "sufficiently diverse corpus" ("corpus assez disparate", Krier 1976: 44) consisting of texts from various genres – poetry extracts, a novel, the Gospels, and newspaper articles covering the society and the politics (for details, see Krier 1976: 45) – and then selecting a sample of "about fifty" ("une cinquantaine", Krier 1976: 44) utterances from each to analyze their syntax. However, her ultimate goal was to analyze the Italian in-

fluence on Maltese and so she excluded all sentences that were deemed entirely Arabic ("les combinaisons de toute évidence arabes", Krier 1976: 45), tailoring her sample to her research question, but making it too biased for other purposes. The sentences in the reduced sample were transcribed and then contrasted with Tunisian Arabic and Tripoli Arabic by means of a consultation with native speakers of those varieties – who were "of the same intellectual level as the authors of those texts" ("de même niveau intellectuel que les auteurs des textes") – in order to determine the degree of influence of Italian (Krier 1976: 45-46). No data on the size of the final corpus is provided, save for the total count of borrowed tokens ("unités de première articulation") at 788 (Krier 1976: 47).

In the analysis of the structure of these utterances, Krier establishes three types of constituents: the predicate ("le prédicat", Krier 1976: 66), the subject ("le sujet", Krier 1976: 68) and the expansion ("l'expansion", Krier 1976: 47). This is defined as "everything that can be joined to the interior of the frame consisting of the minimum utterance without changing the relationships between the constituent elements of this minimum utterance" ("tout ce qui peut être ajouté à l'intérieur du cadre constitué par l'énoncé minimum sans changer les rapports entre les éléments constitutifs de cet énoncé minimum") and can consist of a single moneme, a syntagma or a predicative syntagma (Krier 1976: 73). There are two types of expansions: primary, which expand the sentence; and secondary, which expand one of its composite parts ("un de ses termes", Krier 1976: 76). The constituent order of a Maltese sentence is therefore described in the terms of the ordering of S(ubject), P(redicate) and E(xpansion) and the conclusions Krier reaches are laid out in chapter III, section B, subsection 1.4 (Krier 1976: 79). I cite here the entire section in both the original French and my English translation:

> La séquence essentielle est S+P+Expansion(s). Mais nous avons également relevé les séquences suivantes :
>
> P + S (+E)
> S + E + P (+E)
> E + S + P (+E)
> E + P + S (+E)
>
> Cette liberté de position est due à la mise en valeur stylistique, elle n'affecte en rien l'expérience à comuniquer.
> Quelquefois l'antéposition par rapport au sujet marque l'insistance come dans la phrase nominale :
>
> E2 /a'li:kom 'da:n il ,kmanda'ment 'intom il ʔassi'si:n/
> "pour vous ce le commandement, vous les prêtres",
>
> où la redondance du signifiant du "bénéficiaire" accentue encore cette insistance.
>
> The default order is S+P+Expansion(s), but we have also encountered the following orders:

P + S (+E)
S + E + P (+E)
E + S + P (+E)
E + P + S (+E)

This variation is due to stylistic emphasis and does in no way affect the experience to be communicated.
Sometimes the anteposition with regard to the subject marks the insistence as in the following noun phrase:

E2 /a'li:kom 'da:n il ,kmanda'ment 'intom il ʔassi'si:n/
"for you this commandment, you the priests",

where the redundant reference to the "beneficiary" underscores this insistence.

3.6.2 Summary and evaluation

In broad strokes, Krier joins ranks with Aquilina against Sutcliffe and Vella in describing the default constituent order of Maltese as SV(O) while also noting a great deal of variation and for the first time, explicitly listing specific configurations encountered. In the explanation of said variation (i.e. a description of its regularities), however, Krier once again fails to go beyond a vague reference to "valeur stylistique".

3.7 Kalmár and Agius 1983

3.7.1 Overview

Ivan Kalmár's and Dionisius Agius' 1983 paper on "Verb-Subject Order and Communicative Dynamism in Maltese" (an expanded version of their 1981 paper titled "Verb-Subject Order in Maltese") is the first and so far the only study of Maltese constituent order explicitly framed in terms of linguistic functionalism. Referring to Jan Firbas' work on Functional Sentence Perspective (as evident from the inclusion of the term Communicative Dynamism in the title, see Chapter 2, section 2.4.1), the authors argue that "every proposition in language consists of a link and an advance: the link anchoring the proposition to what has already been communicated (or is known anyway) and the advance representing new or at least unpredictable information" (Kalmár and Agius 1983: 335). Typically, the subject of the sentence acts as the link, but other constituents can also serve as such; those scenarios are the focus of the paper.

Kalmár and Agius accept as a given that the default constituent order in Maltese is SV (Kalmár and Agius 1983: 336, 337), but only if the subject is the link. The aim of their paper is to show that if other constituents serve that communicative function, the VS order is possible and even may be required. To that end, Kalmár and Agius assem-

bled a collection of texts ("four spoken texts, a collection of folk tales, three newspaper articles and some elicited material", Kalmár and Agius 1983: 336) and surveyed various constituent order configurations where VS is possible or even mandatory. Their findings can be summarized as follows (examples are cited with their original numbers and in their original transcription, but without glossing and sources):

I. (O)VS is mandatory if O is the link (Kalmár and Agius 1983: 336):
   (6). *kɔmpjuter ma tri:du:ʃ.*
   'A computer you don't want'.

II. VS is possible in relative clauses modifying the object (Kalmár and Agius 1983: 338, emphasis added):
   (8). *ɛk: ukɔll: fah:rɛt ilpas:i kba:r ʔil ʔúd:i:m li ġ:mlet ma:lta fdan is:et:ú:r*
   'And so it [a visiting Chinese delegation] praised the great steps forward that Malta undertook in this sector'.
   If, however, the verb is the link (and thus S is the new/unpredictable information), the order of constituents in the relative clause is reversed (Kalmár and Agius 1983: 338, emphasis added):
   (11). *ɛk: ukɔll: fah:rɛt ilpas:i kba:r ʔil ʔúd:i:m li ma:lta ġ:mlet fdan is:et:ú:r*

III. AVS is possible if A is the link:
   (15). *il:ejla mí:tet ilmara*
   'Tonight my wife died'.
   In adverbial clauses, so Kalmár and Agius (1983: 340), "the link is the subordinating conjunction" and this is why it comes first; the order of constituents still follows the rules of ordering of link and advance outlined above.

IV. VS is mandatory if the subject is indefinite (Kalmár and Agius 1983: 342):
   (28). *ɛm: ek: jahdem rá:dʒel*
   'A man is working there'

V. VS is mandatory in clauses with *hemm* (and the optional auxiliary *kien*) as the predicate (Kalmár and Agius 1983: 342-343)
   (37). *iku:n ɛm: ilbej:ʝ:a*
   'There will be vendors'.

VI. VS is mandatory in predicates which indicate the emergence of a subject (Kalmár and Agius 1983: 343-344):
   (41). *tela rá:dʒel dʒdí:d*
   'A new (type of) man was born.'

VII. VS is possible in predicates which indicate the appearance of a contextually independent (i.e. new) subject (Kalmár and Agius 1983: 344):
   (42). *u dʒew linglí:zi*
   'The English arrived'.

VIII. VS is possible in clauses which introduce quoted speech (Kalmár and Agius 1983: 344):
   (46). *ʔa:l lil dik ilmara ilhal:í:l*

'Said the thief to the woman.'

3.7.2 Summary and evaluation

In evaluating Kalmár's and Agius work, one must first take note of their method: like Krier before them, they also employ a proto-corpus approach. Unlike Krier, however, they do spare a thought on what corpus linguistics terms accountability (Kalmár and Agius 1983: 337):

> In reporting findings based on the analysis of whole texts one often has to deal with the problem of how to present the data within the space of an article. Clearly there is no room to include entire texts; one can only communicate crucial examples. This of course makes it impossible for the reader to fully reexamine the author's basic data.

Their solution is to provide the exact source of example sentences taken from printed materials, thus enabling readers with access to Maltese-language materials to see for themselves. For the spoken texts, however, the authors rely on "the reader's goodwill to accept our rendering of the context as reliable" (Kalmár and Agius 1983: 337). That, along with the lack of any basic statistical information on their data and the fact that one of the authors (Agius) is also one of the informants for the spoken texts, renders Kalmár and Agius' concern for accountability moot and their results somewhat suspect.

In spite of the methodological issues, Kalmár and Agius' analysis is a valuable contribution to the study of constituent order in Maltese, if only because it is the first attempt to not only map the observed variation, but also to give an explanation of the same. The terminology they use may not be immediately familiar, but the concepts behind it are: what they refer to as "link" and "advance" are nothing but synonyms for "theme" and "rheme" (see Chapter 2, section 2.4.1). Both authors are aware of the alternative terminology – in fact, they specifically avoid using the term "theme" due to Halliday's use of the term in the structural sense (Kalmár and Agius 1983: 336, footnote 2) – so the one they use, they use by deliberate choice. Regardless of metalanguage, the content is clear: with the description of default constituent order a SV and the focus on deviations from it based on the constituent's function as a topic, their work is the first systematic analysis of information structure and its role in Maltese.

In that sense, Kalmár and Agius' paper does partially show what it sets out to show, i.e. that the order of subject and verb in Maltese is to some extent dependent on the pragmatics of the sentence ("If [a constituent other than the subject] serves as link, verb-subject order is possible and in some cases even obligatory", Kalmár and Agius 1983: 336). The findings they present, however, paint a more complicated picture.

Firstly, while they do find that VS order is pragmatically determined, that is only true of 4 out of the 8 instances given, i.e. findings I, II, III and VII above. Here Kalmár and Agius describe what subsequent works (e.g. Borg and Azzopardi-Alexander 1997: 124-128) refer to as the topicalization of objects (finding I) and adverbs (finding III)

in all types of clauses and of subjects in relative clauses (finding II). Pace Borg and Azzopardi-Alexander (1997), topicalization here refers to the constituent moving from its default position to the beginning of the clause, the difference here being that according to Kalmár and Agius, for objects and adverbs in all types of clauses, their default position is before the verb (I and III), while for subjects in relative clauses, their default position is after the verb (II). In the latter case, Kalmár and Agius thus establish that the VS constituent order in relative clauses is the syntactic default and the SV variation is pragmatically dependent. As with Vella (1831), this generalization is confined to one specific type of relative clauses, those modifying the object.

Their findings V concerning existential predicates, VI concerning the emergence of a subject and VII concerning the introduction of a contextually independent subject nearly perfectly describe thetic sentences. VI and VII describe essentially the same scenario, emergence of a new subject, except in VI, it is absolute while in VII, it is relative to the situation. Moreover, the example Kalmár and Agius give for finding VII involves not only an introduction of a new entity, but also a verb of motion. This is – as noted in Chapter 2, section 2.4.4 – a typical example of a thetic sentence.

The sole remaining finding, IV, then stands out as it entails a semantic, rather than pragmatic or syntactic restriction. Moreover, there exist examples to the contrary, such as (1) below.

(1) *Hemm hekk għadd ta' persuni wkoll inġabru biex*
there thus number GEN person-PL also they are assembled in order to
jaraw il- korteo għaddej.
they see DEF procession passing.
'There a number of people assembled to watch the procession pass by.'

Whether this example satisfies the definiteness restriction and thus invalidates finding IV is far from clear, but it only underscores the very preliminary nature of Kalmár and Agius's generalizations concerning the constituent order in Maltese and its pragmatic variation.

Despite the shortcomings cited above, Kalmár and Agius make an invaluable contribution to the study of Maltese constituent order by analyzing it from the point of view of information structure, clarifying for the first time the role information structure plays in the variation of the position of constituents with regard to one another. Their work, while methodologically suspect, thus set the stage for further research into topicalization in Maltese (Fabri and Borg 2002, Borg and Azzopardi-Alexander 2009). Additionally, they have shed more light on related phenomena, such as the VS order in main clauses in narratives already highlighted by Aquilina (1959: 341) and pointed to several phenomena that still remain unaccounted for in Maltese, such as existential sentences.

3.8 Fabri 1993

3.8.1 Overview

Fabri's 1993 *Kongruenz und die Grammatik des Maltesischen* is a generativist analysis of agreement in Maltese based on the formalism of generalized phrase structure grammar (GPSG) and head-driven phrase structure grammar (HPSG) (Fabri 1993: 1-2) and a version of the $\bar{X}$ theory as the mechanism for building the syntactic structures (Fabri 1993: 8).

In the context of his thorough analysis of agreement, Fabri also examines a number of related phenomena in some detail, including constituent order. In general terms, he notes that Maltese constituent order is "relatively free" ("relativ freie Worststellung", Fabri 1993: 7, 131). In terms of his framework-bound analysis, however, Fabri describes Maltese as a configurational language (Fabri 1993: 140, translation mine; see also chapter 2, section 2.3.1):

> Die Beispiele … zeigen, daß die Sb-Phrase im unmarkierten Fall nicht zwischen dem Verb und seinem internen Argument vorkommen darf. Das deutet darauf hin, daß Maltesisch trotz seiner relativ freien Wortstellung doch eine konfigurationale Sprache ist, in dem Sinne, daß es eine strukturelle VP besitzt und keine flache Struktur hat.
>
> Examples … show that the Subject phrase in unmarked cases cannot appear between the verb and its internal argument. This indicates that despite its relatively free word order, Maltese is still a configurational language in the sense that it possesses a structural Verbal Phrase and does not have flat structure.

Having established this, Fabri examines in detail the position of subject NPs ("Sb-Phrase"; henceforth: S) and object NPs (or lexical objects, in Fabri's terminology "Ob-Phrase"; henceforth: O). He does both in syntactic terms, as well as with regard to their role in the information structure of a sentence as topic ("given information", "gegebene Information"; Fabri 1993: 134), focus ("new information", "neue Information"; Fabri 1993: 134) or contrastive topic or focus ("contrasted information", "hervorgehobene Information"; Fabri 1993: 134).

For subjects of intransitive verbs, Fabri finds that if S is the topic, the order of S and V is "completely free" ("völlig frei", Fabri 1993: 137) and the same applies to the position of adverbs with regard to S (Fabri 1993: 138). With S in focus, the situation is much more complex: Fabri notes that acceptability judgments differ based on semantic criteria such as definiteness and animacy or even on verb type (Fabri 1993: 138) and concludes that the full account of the variation is "beyond the scope of this work" ("sprengt den Rahmen dieser Arbeit", Fabri 1993: 139). With transitive verbs, Fabri finds the position of S as the topic with regard to V is also somewhat free and only the VSO order is disallowed (Fabri 1993: 140). In contrast, when a transitive verb has a subject in focus, the subject can only appear left of the verb (Fabri 1993: 141). Fabri, however, goes on to note that a S in focus can appear right of the transitive verb if the

verb bears the direct object clitic in addition to the lexical object (Fabri 1993: 141) and in such situations, the VOS order is in fact the only one that is not permitted. Fabri therefore concludes that in Maltese sentences, there is no "obligatory syntactic slot for the subject" ("Es gibt im Maltesischen keine feste syntaktische Subjekt-Position", Fabri 1993: 142) and the position of the subject is determined pragmatically or semantically (Fabri 1993: 142).

As for objects, Fabri finds that the position of O depends on the presence or absence of object clitics. When object clitics are absent and S is topic, the only configuration that is disallowed is VSO. With (direct) object clitics present, all six configurations are permitted "without a contrastive reading" ("und es gibt keine kontrastive Lesart", Fabri 1993: 144) with adverbials taking any position. In Fabri's interpretation, this confirms the status of the object clitic as the verbal argument which satisfies the verb's θ role and takes the default post-verbal position (Fabri 1993: 144, 146). The role of O (the lexical object) is then solely a pragmatic one: with the clitic present, O is the topic; without the clitic, O is in focus (Fabri 1993: 145-146).

3.8.2 Summary and evaluation

In general terms, Fabri 1993 describes constituent order in Maltese as free (while describing some limitations) with considerable variation and provides not only a detailed description of said variation and its extent, but also attempts to provide generalizations for the same in terms of information structure. Fabri thus in effect argues against any idea of basic or default constituent order in Maltese and makes a case for pragmatic (information structure) considerations being the primary determinant of the ordering of subject, verb and object in Maltese.

3.9 Borg and Azzopardi-Alexander 1997

3.9.1 Overview

Borg and Azzopardi-Alexander's 1997 descriptive grammar of Maltese is the most comprehensive description of Maltese grammar to date and thus the standard work on the subject. Its only disadvantage is its atypical format: as a part of the Routledge "Descriptive Grammar" series, it adheres to the format of the series which is a very detailed typological questionnaire (Comrie and Smith 1977). For some areas of interest, this format makes it somewhat difficult to obtain a big picture of a particular phenomenon, as the relevant information is scattered all over the volume. This is doubly true of constituent order which is referenced in at least 8 different sections. The chief among them is section 1.2.1.2.6. which asks for a description of "the order of the constituents for the

combination of verb, subject, and direct object" and where the following is provided as the answer (Borg and Azzopardi-Alexander 1997: 57):

> The neutral order is Subject-Verb-Direct Object-Indirect Object. Adverbial expressions come last with Manner preceding Place and Time. Variations from this order correspond to specific communicative intentions.

The primary of those communicative intentions is topicalization, discussed at length in section 1.12.1.2.1 which describes the topicalization of objects by movement from their default position to the beginning of the sentence (Borg and Azzopardi-Alexander 1997: 124-125), producing OVS order (no explicit mention is made of the possibility of OSV). In this context, they note that "the SVO order is neutral" and that "NP V NP sequences are never ambiguous between an SVO and OVS" (Borg and Azzopardi-Alexander 1997: 138). As a general remark, Borg and Azzopardi-Alexander observe that "topicalization of the direct object (as well as of the indirect object ...) is such a wide spread characteristic of Maltese, that it even features in Maltese English" (Borg and Azzopardi-Alexander 1997: 126). The topicalization of subjects is also described (Borg and Azzopardi-Alexander 1997: 129), but this only involves suprasegmental phonology (e.g. a phonological break) and encliticization of the verb without any change to the position of the subject with regard to the verb.

In addition to these general considerations, Borg and Azzopardi-Alexander provide a description of constituent order in several types of clauses. For example, they note that in copular clauses with a nominal complement, "the neutral order" is Subject-Predicate, but it can be reversed depending on communicative needs (Borg and Azzopardi-Alexander 1997: 50) and is typically accompanied by specific suprasegmental emphasis (Borg and Azzopardi-Alexander 1997: 118). In case of noun clauses (section 1.1.2.2.2.), they describe VS as "the most neutral" order (Borg and Azzopardi-Alexander 1997: 33) while SV, "with an accompanying suprasegmental change on the superordinate verbal expression", is used to "emphasize the superordinate verb" (Borg and Azzopardi-Alexander 1997: 33). For a subtype of noun clauses, the so-called adjectivalized noun clauses, they describe "[t]he relative order of the arguments in the adjectivalized noun clause" as SVO (Borg and Azzopardi-Alexander 1997: 34-35). They also describe the reversal of default SV to VS in comparative and equative adverbial clauses (Borg and Azzopardi-Alexander 1997: 45-46), various types of emphasis through clefting (Borg and Azzopardi-Alexander 1997: 117-121) and provide a detailed account of constituent order variation in questions (Borg and Azzopardi-Alexander 1997: 3-27).

3.9.2 Summary and evaluation

The primary shortcoming of Borg and Azzopardi-Alexander 1997, one that in no way distracts from the fact that this is the most complete description of Maltese to date, is the idiosyncratic format. This, along with its broad focus, results in a number of

incomplete generalizations and curious omissions, the chief among them the lack of any mention of focus articulation in Maltese, save perhaps for references to suprasegmental realization thereof. More importantly, the generalizations provided by Borg and Azzopardi-Alexander regarding constituent order are subordinate to the methods and goals of the questionnaire. And so while Borg and Azzopardi-Alexander seem to contradict Fabri's (1993) conclusions and identify SVO as the unmarked ("neutral") constituent order, ascribing variation to movements (primarily topicalization), this may be merely an artifact of the format and not the result of their adherence to a particular framework (i.e. generative grammar) to the theory of which concepts like "movement" belong.

3.10 Fabri and Borg 2002

3.10.1 Overview

In 2002, Fabri and Borg conducted a study which functions as a follow-up to and refinement of earlier analyses of constituent order and the role of information structure in it (primarily Fabri 1993 and Borg and Azzopardi-Alexander 1997). The stated primary purpose of the paper is typological: Fabri and Borg set out to apply Greenberg's correlations to Maltese, but noting that it can be difficult to identify basic order in "so-called structurally non-configurational languages like Maltese, i.e. languages with a certain degree of free word order" (Fabri and Borg 2002: 354). Working with the hypothesis that "Standard Maltese is a discourse configurational language" (citing Kiss 1995a, see Chapter 2, section 2.3.4), their aim is:

(a) "to determine whether one basic or unmarked word order exists, and
(b) to give structural and functional explanations for the other 'derived' word orders" (Fabri and Borg 2002: 355).

In actuality, the paper is almost entirely focused on the latter goal and is thus a thorough investigation of information structure – or rather the possible articulations of topic and focus – in Maltese "simple mono-clausal construction[s] ... with simple intransitive and mono-transitive verbs and ... excluding ditransitive and complex verbs, as well as adverbials and subordinate clauses" (Fabri and Borg 2002: 355). In lieu of defining topic and focus in one of the traditional ways (old vs. new etc.), Fabri and Borg wisely adopt a practical solution in the form of a set of questions which serve to define the discourse context and thus elicit the desired articulation in terms of information structure (Fabri and Borg 2002: 355). Taking into account two additional parameters, encliticization and stress, they construct a set of questions with respective answers modified based on the investigated parameters (constituent order, stress, clitics) and then judge their felicitousness (Fabri and Borg 2002: 358).

The major conclusion of this investigation is that while in Maltese, there is a considerable degree of variation, the "unmarked order in a sentence without a clitic on the verb" is:

(a) "SV and VS with stress on V for intransitives
(b) SVO and OVS with stress on O for transitives" (Fabri and Borg 2002: 362)

In terms of other factors influencing constituent order in Maltese, Fabri and Borg find that stress interacts with information structure to a larger extent than previously thought and so NPs in focus are always stressed, while topic NPs never are (Fabri and Borg 2002: 362). They also confirm Fabri's (1993: 144-146) observation that topic NPs always require the presence of a co-referential clitic and provide a list of possible configurations for specific pragmatic functions and draw generalizations from a subset of them. By and large, however, a detailed and comprehensive description of information structure and its role in Maltese constituent order variation is beyond the scope of their work.

3.10.2 Summary and evaluation

Fabri and Borg 2002 remains to this day the most complete study of constituent order and information structure in Maltese. This is despite the fact that some of the findings were refined by further research, such as the generalization that "pronominal clitics can only be coreferential with definite NPs" (Fabri and Borg 2002: 360) which has subsequently been shown to be incorrect (Fabri 2011, Čéplö 2014: 206-208) or the description of OVS without clitics on the verb as the only possible construction for object in focus which I called into question in the course of my analysis of object reduplication and related phenomena in Maltese (Čéplö 2014: 212-213).

The chief shortcoming of Fabri and Borg 2002 is their methodology: first, they narrow the scope of their research to "basic constructions" (Fabri and Borg 2002: 355) and secondly, they do not provide any definition of the term "basic order" or "unmarked"; consequently, it is unclear what those terms refer to. That their findings are based on introspection (structured and detailed though it is) only underscores the fact the results of their investigation are, as they themselves admit (Fabri and Borg 2002: 362), tentative.

3.11 Other

In the last decade, a number of works appeared which touch upon the issue of constituent order in Maltese, some of them expanding on previous works, others treating the subject from a typological standpoint. The former group includes a paper read by

Borg and Azzopardi-Alexander at the first meeting of *Għaqda Internazzjonali ta' Lingwistika Maltija* (Borg and Azzopardi-Alexander 2009) which discusses topicalization in Maltese, a subject to which devoted much attention in their 1997 grammar. This paper expands on that description by confirming their view of topicalization as a syntactic movement (Borg and Azzopardi-Alexander 2009: 79), but again without declaring allegiance to a particular framework. Primarily, however, this paper offers a detailed analysis of the suprasegmental component of topicalization, as well as points out the fact that Maltese allows multiple topics (cf. my follow-up analysis of Hanging Topic Construction and Clitic Left Dislocation in Čéplö 2014: 205-212).

The latter group, typological descriptions of Maltese constituent order, includes an overview of Maltese as one of the new languages of the European Union (Fabri 2010). In this paper, Fabri describes Maltese as "a topic-oriented language, especially in the spoken form" remarking that "apart from the subject noun phrase, all kinds of object phrases can be placed at the beginning of the sentence" (Fabri 2010: 793). Fabri then describes the constituent order of Maltese as "relatively free", essentially summarizing his findings from Fabri 1993 (see above) and concluding that SVO is the unmarked order "with the other variants being used mainly contrastively, given the appropriate intonation" (Fabri 2010: 793-794).

And finally, in the course of their analysis of Maltese complement clauses, Borg and Fabri (2016) describe Maltese as "a discourse configurational ... language, especially in its spoken form". They go on to note (echoing previous works by Borg, Azzopardi-Alexander and Fabri) that "most constituents of the sentence can be topicalized by being placed at the beginning of the sentence" and to describe the phonological aspects of such constructions (Borg and Fabri 2016: 417).

3.12 Conclusion

As this discussion has shown, two constant themes are interwoven throughout the history of the study of Maltese constituent order: first, there is the question of what is the default (unmarked, basic, dominant) constituent order in Maltese. This has been answered in at least two different ways: verb-first, as argued by Sutcliffe 1936 and Vella 1970; or subject-first, as described by Aquilina 1959, Kalmár and Agius 1983, Borg and Azzopardi-Alexander and others. The other theme is that of classifying Maltese constituent order as "free" (e.g. Fabri 1993: 7, 131 and Fabri 2010: 793), including synonyms like "discourse-configurational" (Fabri and Borg 2002, Borg and Fabri 2016) and "topic-oriented" (Fabri 2010: 793, Fabri and Borg 2017: 83), all of which describe Maltese as a language where "constituent order, at sentence level is strongly influenced by pragmatic factors, in particular topic and focus, contrast and emphasis, more than by syntactic factors" (Fabri and Borg 2017: 83). In this context, a number of authors note a great deal of variation in Maltese constituent order (Sutcliffe 1936: 211, Krier 1976:

79, Fabri and Borg 2002) and attempt to account for it (Borg and Azzopardi-Alexander 1997, Fabri and Borg 2002).

Both these analyses can be shown to have serious shortcomings: for the question of the default (unmarked, basic, dominant), the chief one is obviously the lack of general agreement. Additionally, there are multiple methodological issues, ranging from the lack of a meaningful definition of "default (unmarked, basic, dominant)" constituent order, through the lack of detailed studies on clause-type level (with Borg and Azzopardi-Alexander 1997 as sole attempt to do so in a systematic manner), all the way to the fact that most such studies have been introspective at best, impressionistic at worst. Even those that employed some sort of empirical approach (Krier 1976, Kalmár and Agius 1983) did so more than imperfectly, rendering their conclusions tentative at best. Much of this also applies to works which describe Maltese constituent order as free or pragmatically determined, which additionally have problems of their own. And so for example even those studies that provide a detailed account of the possible variation based on pragmatic (information structure) factors (Borg and Azzopardi-Alexander 1997, 2009; Fabri and Borg 2002) essentially only described potentiality, i.e what options are available to speakers of Maltese, but did not (except in the broadest terms, e.g. Borg and Azzopardi-Alexander 1997: 126) provide a description of how those possibilities are instantiated.

In what follows, I set out to remedy these shortcomings.

4 Interlude: Research questions

4.1 Introduction

Having established the terminology and general aproach used in this thesis (chapter 1) and its context (chapter 2) and surveyed the field (chapter 3), I will now proceed to the titular subject of this work, constituent order in Maltese. In this interlude, I will narrow down the focus by setting research questions to provide clear and achievable goals for the research herein.

4.2 Research questions

This work seeks to provide the answers to the following questions:

1. What is the dominant constituent order in Maltese?
2. What is the variation in dominant constituent order in Maltese?
3. What are the deviations from the dominant constituent order in Maltese?
4. What are the determinants of variation in Maltese constituent order?

4.2.1 Research Question 1: What is the dominant constituent order in Maltese?

4.2.1.1 Question in context

In addition to general descriptive considerations, Research Question 1 is motivated by previous research into Maltese constituent order, most of which assumes (to some extent) the existence of default (unmarked, basic, dominant) order (see Chapters 2 and 3). Consequently, the main task at hand is to determine whether there is a default (unmarked, basic, dominant) constituent order configuration in Maltese and what it is. The purpose of this questions is therefore to check previous work on the subject and settle the issue of whether the default (unmarked, basic, dominant) constituent order is verb-first, as some (Sutcliffe 1936, Vella 1970) would have it or whether its subject-first, as others (Aquilina 1959, Kalmár and Agius 1983, Borg and Azzopardi-Alexander 1997) argue; this question thus addresses the typology of Maltese. In this context, several other claims have been made regarding the nature of Maltese constituent order, most notably the description of Maltese as a discourse-configurational language (e.g. Fabri and Borg 2002 and Borg and Fabri 2016) or a language having a free word order (Fabri 1993: 131, Fabri 2010: 793), and I will endeavor to address those as well.

In light of the typological nature of this investigation, constituent order in Maltese will be primarily analyzed in terms of Dryer's SV/VS and VO/OV typology (Dryer 1997, Dryer 2013b). The reasons for this are two: first, as Dryer notes, such typology "is based

not only on clauses containing both a nominal subject and a nominal object but also on clauses containing just one of these" (Dryer 2013b: 269). This is not only appropriate in light of the existence of transitive and copular clauses and the fact that clauses featuring only one of the core verbal arguments "occur much more frequently" (Dryer 1997: 70), but it is also particularly relevant for languages like Maltese where the nominal subject is not obligatory in verbal clauses. Secondly, the binary typology allows for fine-grained analysis and better visualization, especially when multiple objects of analysis (i.e. various types of clauses) are involved. Nevertheless, the six-way Greenbergian typology will be occasionally referred to for illustration and for comparison with previous studies on Maltese constituent order.

The concept of "dominant constituent order" is likewise borrowed from Dryer (2013a). As cited in chapter 2, section 2.2.1.3, Dryer defines the dominant order (whether constituent order or word order) as follows:

> The rule of thumb employed is that if text counts reveal one order of a pair of elements to be more than twice as common as the other order, then that order is considered dominant[.]

This definition is clear and empirically based, explicitly referring to language corpora statistics ("text counts"); as such, it is perfectly suitable for the approach used in this work as outlined in Chapter 1, section 1.2. As this definition of dominant constituent order is also the one used in a major work on language typology (*The World Atlas of Language Structures*), the data gained in answering the question can be used to supplement and/or correct the information provided for Maltese therein or any other work on comparative typology.

4.2.1.2 How to answer it?

To answer Research Question 1, I will examine a syntactically annotated corpus (treebank) of Maltese (see section 4.3.1 below and Chapter 6 for details) to determine the distribution of SV/VS and VO/OV orders in clauses contained therein. As noted in Chapter 1, section 1.3.2.2, the syntactic annotation used in the treebank is based on the UD standard. Accordingly, the quantitative analysis of constituent order configurations will be performed not only across all clauses, but also separately for main clauses and various types of dependent clauses as defined by UD (see Chapter 6, section 6.4).

This type of analysis, along with the concept of dominant constituent order used here, necessitates establishing and defining two types of variation from dominant constituent order: the first, termed "variation" proper,[1] will be used for situations where one or several types of clauses (however defined) display dominant constituent order different from that in other clause types or across all clauses. The second term, "deviation", will be used for the non-dominant configuration: recall that the dominant con-

1 This is essentially equivalent to Bakker's "flexibility" (Bakker 1998: 387).

stituent order is defined here in statistical terms as the configuration which is twice as frequent as the other option; that other option will then be referred to as "deviation" or "deviant order".

4.2.2 Research Question 2: What is the variation in dominant constituent order in Maltese?

4.2.2.1 Question in context

This question seeks to address the second part of the assumption underlying Research Question 1, i.e. the existence of alternative dominant constituent order(s) in certain types of clauses, including situations when the dominant one cannot be established. As such, this question builds on previous works which identify SV as the default, but noted the possibility of VS in some types of relative clauses (starting with Vella 1831: 255) and especially on Kalmár and Agius (1983) who identified a number of contexts where VS appears to be mandatory.

4.2.2.2 How to answer it?

The answer to Research Question 2 will be provided using the same type of quantitative analysis employed in answering Research Question 1, as a complement to it.

For cases where the dominant constituent order cannot be determined, i.e. the ratio between SV/VS or VO/OV is lower than 2:1 for either of the options, answering this question will involve a detailed analysis of the determinants of said variation which will be addressed in Research Question 4.

4.2.3 Research Question 3: What are the deviations from the dominant constituent order in Maltese?

4.2.3.1 Question in context

This question focuses on contexts where the dominant constituent order could be established, but the clauses in question exhibit the non-dominant order. The primary purpose of this question is to check a number of observations made regarding the topicalization of the direct and indirect object (see Chapter 3), best summarized by Borg and Azzopardi-Alexander who describe it as "such a wide spread characteristic of Maltese, that it even features in Maltese English" (Borg and Azzopardi-Alexander 1997: 126).

4.2.3.2 How to answer it?

The answer to this question is provided as a complement to the answer to Research Question 1.

4.2.4 Research Question 4: What are the factors that cause variation in dominant constituent order?

4.2.4.1 Question in context

Having established what the variation in dominant constituent order is, I will focus on those types of clauses that exhibit said variation, analyzing their structure and attempting to determine what causes said variation.

The issue of to what extent is the variation (and deviation) in constituent order a phenomenon rooted in grammar (i.e. syntax) and to what extent it is a pragmatic (information structure) phenomenon is one of the major problems in the study of constituent order. That both are involved is now taken for granted: schools of thought that started out arguing for the former now recognize the role of pragmatics and included information structure concepts in their theory (e.g. generative linguistics and the concept of discourse-configurationality or the Cartographic Project, see Chapter 2, section 2.3); those schools of thought that do focus on the role pragmatics have always explicitly recognized the role that syntactic constraints play in constituent order variation (e.g. Firbas 1992: 118, see Chapter 2, section 2.4).

Consequently, both approaches have postulated their own version of a list of rules that contribute to the instantiation of a particular configuration: in FSP, these include at the very least the principle of grammatical function, principle of coherence of members, the principle of FSP and the principle of sentence rhythm (Firbas 1992: 117, Mathesius 1961: 180-191). In the generative tradition, one study arrives at the following list (Siewierska 1988: 263):

(a) grouping relations
(b) grammatical relations
(c) thematic relations
(d) semantic roles
(e) syntactic features (e.g. categorial status, internal categorial structure, tense, aspect, modality, mood, finiteness etc.)
(f) semantic features (e.g. animacy, humanness, definiteness, referentiality, etc.)
(g) pragmatic factors (e.g. perceptions of salience or dominance, familiarity, iconicity, relative identifiability etc.).

This list (along with similar inventories provided by, say, the Cartographic Project) is long and the issues involved are complex; as such, they cannot be given full consideration in a work like this one. In my efforts to map and account for constituent order variation and deviation in Maltese, I will therefore focus on two major players: syntax and information structure.

The former is obvious and in line with the descriptive approach employed here; same goes for the practical aspect of it and so in accordance with the definition of syntax in Chapter 1, section 1.3.3.2, I will examine the syntactic factors in terms of dependency relations. Additionally, however, I will introduce into the discussion concepts relating to description of language in quantitative terms, some relatively straightforward, like

clause length (cf. Köhler 2012: 142-146), some less so, like "heaviness" (Arnold et al. 2000) which refers to the structural complexity of a constituent (i.e. in our case, the length of the catena which has a core argument of the predicate as the head) and which has repeatedly been found to influence the ordering of constituents (Arnold et al. 2001: 51, Stolz 2011 for Maltese).

As for information structure, this will be primarily discussed in terms of describing constituent order variation, as well as while addressing previous typological classifications of Maltese (see Chapter 2). The analysis of constituent order deviation (assuming any is found) where information structure plays a role is, for the most part, outside of the scope of this work.

4.2.4.2 How to answer it?

The clauses which exhibit variation in dominant constituent order will be analyzed either computationally or manually, depending on the number of clauses involved. I will examine their syntactic, semantic and pragmatic properties, primarily as compared to those clauses that exhibit the dominant order.

4.3 Data and methodology

4.3.1 Data

The analysis as outlined above will be performed using corpus data. These come in two forms: the primary source of data for the quantitative analyses will be the Maltese Universal Dependencies Treebank v1 (MUDTv1). This treebank, annotated according to the Universal Dependencies annotation standard, version 1 (UD v1; Nivre, Ginter et al. 2014; Nivre, de Marneffe et al. 2016), is the first ever compiled for Maltese; I have created it myself for the purpose of this thesis and it will be made available to the public upon its defense with the hopes that should I fail to achieve the goals set herein, the treebank will at least be of some use to someone. Chapter 6 describes in detail the composition of the treebank, the annotation decisions and the reasoning behind them.

The other data source is the general corpus of Maltese which I have also compiled myself (*bulbulistan maltiv3*, *BCv3*), described in Chapter 5. *BCv3*, despite being only annotated with the bare minimum of linguistic information and thus incapable of serving as data source for the actual analysis of constituent order in Maltese, nevertheless plays a crucial role here: first, it is the primary source of texts for MUDTv1. Secondly, it provides material for the analysis of syntactic phenomena that are being described as a part of the annotation of syntactic relations in MUDTv1 from the fundamentals of linguistic analysis like part-of-speech tags all the way to verbal valency (see Chapter 6, section 6.4.4.2), which is crucial for the phenomena under study. And finally, it will be

used to check and test information obtained from the analysis of MUDTv1. As such, it is an integral part of this work.

4.3.2 Methodology

The primary tools employed here are those of descriptive statistics. Using the data from MUDTv1, I will provide the basic statistics on the distribution of the orders of the subject and the predicate (SV/VS) and the object and the predicate (VO/OV), including visualizations thereof, to determine the dominant constituent order both across the entirety of MUDTv1, as well as per clause type. Wherever applicable, I will also apply methods for the testing of statistical significance, primarily to determine whether the differences encountered (such as the ratio of one configuration versus another) are real or only due to chance. And finally, in scenarios such as those where no dominant order can be established, I will use statistical modeling to account for it.

5 *BCv3*: A corpus of written Maltese

5.1 Introduction

This chapter discusses the general corpus of Maltese which serves the primary source of data for the analysis of Maltese syntax and the preparation of the treebank. The process of data collection and selection, as well as the preparation, processing and annotation of the data are described in detail.

5.2 History of Maltese corpus linguistics

First attempts to collect machine-readable data for Maltese took place in the course of the MaltiLex Project (Rosner et al., 2000, Bovingdon and Dalli 2006). The stated aim of MaltiLex was to construct an electronic Maltese lexicon, based on corpus data; however, the resulting corpus was of a relatively small size and lacked any meaningful structural or grammatical annotation. More recently, Ussishkin, Francom and Woudstra (2009) used web resources to create a medium-sized corpus, primarily for use in the extraction of lexical resources to inform their experimental work on Maltese lexical processing. Two recent European initiatives, Clarin[1] and METANET4U,[2] have also provided impetus for further development of Maltese language resources: while work within Clarin mainly focused on the digitization of resources within the humanities, the METANET initiative aimed to build a common, Europe-wide infrastructure to accommodate corpora and text and speech processing tools.

These efforts culminated in 2011 with the publication of two 100-megaword corpora of modern Maltese, the *Korpus Malti* (a part of the Maltese Language Resource Server)[3] developed by Albert Gatt at the University of Malta and the *bulbulistan* corpus[4] compiled by myself (see Gatt and Čéplö 2013 for a preliminary description of both). Originally conceived as independent projects, their most recent versions (*Korpus Malti v3.0* and *bulbulistan maltiV3*, henceforth *MLRSv3* and *BCv3* respectively) have taken a step towards the eventual integration of both corpora into a single resource by sharing data, adopting standardized processing methods, expanding their reach to over 200 million word tokens and developing a common part-of-speech tagging scheme.

While *MLRSv3* serves as the focal point of Maltese corpus linguistics, *BCv3* continues its existence as a separate entity for technical reasons, legacy reasons and as the data source for a number of special projects, of which this thesis is the primary one.

1 clarin.eu (last consulted on February 28th 2018)
2 metanet4.eu (last consulted on February 28th 2018)
3 mlrs.research.um.edu.mt (last consulted on February 28th 2018)
4 bulbul.sk/bonito2

5.3 Data composition

5.3.1 Data selection

Both *MLRSv3* and *BCv3* are opportunistic corpora by nature; in Malta, there is no legal and logistic infrastructure in place comparable to, say, that by which the Czech National Corpus is provided with data from publishers. The core of both corpora is therefore composed of easily obtainable texts, namely online newspapers, texts produced by the Parliament of Malta and data freely available in digital form (such as Wikipedia entries) which is supplemented with any texts that are available and lend themselves easily to automated processing.

Despite this "beggars can't be choosers" approach to corpus building, a number of limitations was put on the texts included in *BCv3* in line with the focus of this dissertation (see the definitions of Maltese in Chapter 1, section 1.3.2.1). The texts selected for inclusion had to be:

I. Written
   All texts must originate in writing. There are some borderline cases, such as sermons or the transcripts of parliamentary debates which theoretically record speech. They do not, however, capture the distinguishing properties of speech (prosody, interruptions, false starts, turn taking etc.) and furthermore, a comparison of selected transcripts of parliamentary records and the original audio recordings has made it clear that some form of editing or normalization has taken place in the conversion. As such, they can be safely considered as having originated in writing.
II. Original
   All texts must be original compositions by native speakers of Maltese. For this reason, a large corpus of European legislation was excluded from *BCv3*, as was the Bible and works of fiction translated from English, French and Spanish. Some types of texts, like Wikipedia entries or certain portions of the magazine *Lil-Hbiebna*, straddle this fence as a part of them appears to have originated as translation, but the vast majority of texts from both sources can be safely considered original compositions.
III. Public
   Only texts available publicly or intended for public consumption (i.e. unpublished or not-yet-published works by established authors) are included.
IV. Recent
   Only texts that originated within the first two decades of the 21st century were included. This criterion has been somewhat stretched by including a number of texts that are older, such as a few works of fiction dated to the 1990s, two works published in the 1980s and some works of fiction with a publication date after 2000 which are in fact reeditions of earlier works. However, a substantial subcorpus of

pre-1950 fiction was excluded from *BCv3*, as were all available historical materials and parliamentary documents dated before the year 2000.

5.3.2 Text types

The texts collected for *BCv3* fall roughly into four groups based on the origin and/or source. These groups have been defined as text types in corpus metadata as follows:

Text type	Description
newspaper	online newspapers
parliament	records retrieved from the website of the Parliament of Malta
fiction	imaginative literature (excluding poetry) and blogs
non-fiction	academic texts, popular science, sermons, Wikipedia entries

Tab. 5.1: Text types in *BCv3*

For some text types, a subtype could also be determined. This can be either internal, i.e. based on the content and/or the classification provided by the source itself (e.g. local news, international news and sports for newspapers), or again external (novels, short stories and blogs for fiction). Text subtypes will be discussed in the following sections whenever appropriate.

5.3.2.1 Text type: newspaper

The texts that fall under the newspaper text type are almost exclusively source from online editions of Maltese newspapers and dedicated news sites using various web-crawling techniques. Table 5.2 lists these sources along with the dates from which I was able to scrape texts.

Source	Date from	Date to	Tokens
Il-Ġens Illum (C)	2011-04-12	2012-07-23	6,643,087
Illum pre-2016	2006-11-12	2010-05-30	4,320,925
Illum post-2016	2013-11-17	2017-01-12	3,458,919
INewsMalta (L)	2012-07-02	2017-01-31	9,585,674
Kullħadd post-2016 (L)	2016-05-01	2017-04-27	1,031,475
Lil-Ħbiebna (C)	2005	2010	422,178
L-Orizzont (L)	2005	2013	44,925,117
MaltaRightNow (N)	2009-09-03	2015-06-06	17,012,671
NETNews (N)	2014-03-20	2017-04-27	6,978,239
Newsbook	2012-10-10	2017-01-09	8,700,580
It-Torċa (L)	2005	2013	10,041,192
Total			113,120,057

Tab. 5.2: Text type newspaper in *BCv3*

In the interest of enabling the surveying of the politically fractured Maltese society and the effects of this polarization on Maltese, those newspapers and news sites operated by one of the two major political parties are marked as such in the list above, with (L) for the Labor Party and (N) for the Nationalist Party. News sources controlled by the Catholic Church, another major player in Maltese politics, are marked with (C).

For *NET News*, *Newsbook*, *Illum* (pre-2016) and *MaltaRightNow*, a comprehensive list of subtypes was established based on their editorial board's internal classification of individual pieces. The full list of these subtypes is included in Appendix A.

5.3.2.2 Text type: parliament

This text type includes documents produced by the 9th through 12th Legislature of the Parliament of Malta dated from 2000 to 2017. These documents are available to the public on the Parliament's website. Of the various file types, only Word documents (*.doc or *.docx) were selected for inclusion as they contain the majority of the Parliament's document production and can be easily processed automatically.

The parliament texts can be divided into 5 subtypes: meeting minutes, debates, parliamentary questions, agenda and other. The agenda texts have been excluded from consideration due to their repetitive content mostly consisting of lists of dates, events to be held, issues to be discussed and questions to be asked. The basic statistics for the remaining 4 text types can be found in Table 5.3 below.

Subtype	Tokens
debates	75,494,440
minutes	5,896,493
questions	15,487,417
other	1,455,116
Total	98,333,466

Tab. 5.3: Text type parliament in *BCv3*

5.3.2.3 Text type: fiction

The core of this text type consists of works of imaginative fiction (belles lettres), both novels and short stories, published in Malta between 2000 and 2017 supplemented with two novels and a short story collection from the 1980s and 1990s (see Table 5.4). The works included here were obtained in two ways: about a half was scanned from hard copies from my own library, processed with an OCR software and checked for errors. As *BCv3* resides on a Slovak domain on a server physically located in Slovakia, these texts are included here pursuant to section 44 of the Slovak Copyright Act (185/2015) which allows the use of copyrighted material for non-commercial research. The other half of works included in the fiction text type was provided by their authors, either directly or through Merlin Publishers; as such, they are included in *BCv3* with the permission of their respective authors and the publisher. The full list of works is included in Appendix A.

This text type includes a number of blogs written by various Maltese authors which expand on their published works and provide insights into their creative process, as well as texts from the *Għidli Mitejn* project[5] where both established authors and members of the general public can submit their short stories of 200 words or less.

Subtype	Tokens
Novel	1,969,950
Short story	263,548
Blog	204,333
Total	2,437,831

Tab. 5.4: Text type fiction in *BCv3*

The full list of all sources can be found in Appendix A.

5 facebook.com/ghidlimitejn/ (last consulted on February 28th 2018)

5.3.2.4 Text type: non-fiction

The core of this text type consist of four books (all scanned) and the Maltese Wikipedia. The entries from the Maltese Wikipedia were extracted from the Wikimedia dump[6] using the wikiextractor Python script set.[7] For the processing pipeline, each entry was considered a separate document and thus received its own <doc> tag.

And finally, two sets of texts retrieved from the web were included here as well: the first is a collection of literary criticism and reviews by Patrick Sammut published on his blog,[8] the second comprises homilies by Malta's archibishop published on the Church's Maltese website.[9]

Table 5.5 below provides basic information on the non-fiction text type in *BCv3*; the list of the four books included in this text type can be found in Appendix A.

Source	Tokens
Books	277,945
Blog (2010-11-15 – 2016-12-06)	169,500
Sermons (retrieval date 2017-05-06)	147,261
Wikipedia (export date 2017-04-23)	1,757,024
Total	2,351,730

Tab. 5.5: Text type non-fiction in *BCv3*

5.3.2.5 Summary

Table 5.6 below summarizes the composition of *BCv3* by text type. These numbers reflect its opportunistic and imbalanced nature.

Text type	Documents	Sentences	Tokens	%
newspaper			113,120,057	52.31%
parliament			98,333,466	45.47%
fiction			2,437,831	1.13%
non-fiction			2,351,730	1.09%
Total	313,499	9,769,815	216,243,084	100%

Tab. 5.6: Text types in *BCv3*

In the next sections, I will describe the process of preparation and encoding of the corpus, as well as its enrichment with basic linguistic annotation.

6 bit.ly/2CmzEnE (last consulted on February 28th 2018)
7 bit.ly/2F87jCw (last consulted on February 28th 2018)
8 frokna.blogspot.com (last consulted on February 28th 2018)
9 thechurchinmalta.org (last consulted on February 28th 2018)

5.3.3 Data processing

5.3.3.1 Text conversion and cleaning

All texts were first converted to UTF8-encoded text files. For *BCv3*, all texts which are not written in proper Maltese orthography (mostly older editions of *L-Orizzont* and *It-Torċa* and the entire pre-2016 version of *Kullħadd*) were automatically removed from the text pool, as were texts with encoding conversion errors (almost exclusively older parliamentary texts). The remaining files were then processed in a pipeline comprising a text cleaner, a sentence splitter, a language identifier and a tokenizer implemented in a single Perl script.

5.3.3.2 Text cleaning

In the text cleaner, each file was first slurped (i.e. the entire contents were read into a single string). Any existing text division (chapters, unnumbered sections and paragraphs) were removed, converting them to end-of-sentence symbols (EOS) represented by a special character sequence ("__"). The content of the files was then normalized by removing superfluous characters (e.g. double spaces or tabs), converting all hyphen-like characters to hyphen-minus (U+002D) and all double quotes and equivalent characters to typewriter quotes (U+0022). The cleaned text was then passed to the sentence splitter as a Perl string.

5.3.3.3 Sentence splitting

In the sentence splitter, the slurped string was split into sentences by inserting new EOS after every full stop, ellipsis (when followed by a capital), exclamation mark and question mark; these were first joined to the following quotes and square brackets. In technical terms, this was implemented as a regular expression (regex) with a look-ahead followed by another regex-based correction of errors resulting primarily from the use of full stop after abbreviations (e.g. Dr. or St.) and initialisms. This unsophisticated solution was found to be much more effective than learning-based sentence splitting approaches such as the one implemented by NLTK,[10] Apache OpenNLP[11] or LingPipe.[12] The slurped string was then split on EOS and fed into a Perl array of sentences.

The first version of the sentence splitter also treated the colon as sentence-end punctuation, but this behavior was removed in subsequent versions, largely in light of the treatment of clauses separated by a colon in UD (see chapter 6, sections 6.4.4.4.4 and 6.4.4.15.2 on ccomp and parataxis, respectively). The only downside to this solution is apparent in the parliamentary records where particularly in debates, many

10 nltk.org (last consulted on February 28th 2018)
11 opennlp.apache.org (last consulted on February 28th 2018)
12 alias-i.com (last consulted on February 28th 2018)

colon-separated clauses consist merely of the identification of the speaker. They are therefore considered dependents of the main clause in a list relation, same as e.g. chapter numbers or numbered list separators (see Chapter 6, section 6.4.4.15.1); an inelegant, but consistent solution.

5.3.3.4 Tokenization

In both *MRLSv3* and *BCv3*, the tokenization is based on the official orthography of Maltese. This uses the basic Roman alphabet (i.e. the Unicode Basic Latin block) supplemented with vowel characters with grave from the Unicode Latin-1 Supplement block and four Maltese-specific characters: ċ (U+010B) / Ċ (U+010A), ġ (U+0121) / Ġ (U+0120), ħ (U+0127) / Ħ (U+0126) and ż (U+017C) / Ż (U+017B).

Additionally, however, there are two characters which are normally classified as punctuation that should be considered a part of the Maltese alphabet proper. The first of these is the apostrophe which has a special function: it is used to indicate the ellided għ (etymologically [ġ] or [ʕ]) at the end of a word, e.g. *ta'* (< NA *mtāʕ*) or *sema'* (< CA *samiʕa*). The apostrophe is also used to indicate the ellision of a vowel in single syllable prepositions such as *bi*, *fi*, *sa* or (rarely) *ma'* and the reduced form of the negator *ma* when the word they are governed by begins with an accented syllable. And finally, the apostrophe (or one of its visually equivalent alternatives) is also commonly used in place of vowel characters with grave, as in *awtorita'* (instead of *awtorità*) or *ċioe'* (instead of *ċioè*). This practice is in violation of Maltese orthography rules, but it is very common, especially in journalistic texts where the chance for the incorrect version to be used is about one in three (e.g. for *awtorità*, the correct form is used 65,683 times in *BCv3* while the incorrect version *awtorita'* crops up 39,219 times). The second such character is the hyphen (more specifically, hyphen-minus). This is used to join the definite article *il-*, its assimilated forms and its fused forms to the following word, as in *il-gżira* "the island" or *biċ-ċavetta* "with the key".

Consequently, the tokenization in in *BCv3* essentially consists of first separating out all the punctuation (except for hyphens and apostrophes) and then splitting what remains on spaces, hyphens and apostrophes. This solution, like many in NLP, is quick and dirty and works reasonably well.

There are, however, two problems with this approach to tokenization: first, the tokens with fused articles (DEM_DEF, GEN_DEF, LIL_DEF and PREP_DEF) and fused pronouns (GEN_PRON, LIL_PRON and PREP_PRON) are not split into their constituent parts. This is an issue which does not have any major effect on most purposes the corpus can be used for and in any case, it can be easily solved using simple rule-based approach. Secondly and more importantly, this solution does not split off the direct object clitics, indirect object clitics and the negative suffix *-x* from the words they attach to (verbs, pseudoverbs and personal pronouns). This is a violation of the UD guidelines according to which syntactic, not orthographic words are the main unit of annotation (Nivre, Ginter et al. 2014). In Maltese, unfortunately, this would require full morpholog-

ical and syntactic analysis of verbal forms, as clitics are integrated into the structure of the verb and there is some degree of ambiguity (e.g. *ktibtu* can be analyzed both as *ktib-tu* "they/you.PL wrote" as well as *ktib-t-u* "I/you wrote it"). Such an analysis is well beyond the scope of this work – in fact, the incorporation of clitic splitting into automated tokenization has only become possible with recent advances in Maltese computational morphology (Borg 2016). This ommission will therefore be addressed in the next release of both *MLRSv3* and *BCv3*. As for the present work, its primary goal remains unaffected by it: the position of the clitics does not vary and is thus not directly relevant for the analysis herein; in situations where the presence of clitics is of import to the phenomena under discussion, manual analysis will be performed.

In technical terms, tokenization was implemented as a regular expression with a look-ahead for hyphens and apostrophes with subsequent rule-based error correction for abbreviations, contractions, acronyms and other irregular phenomena.

5.3.3.5 Language identification

As the final part of tokenization, a language identification algorithm was implemented to cope with the bilingual nature of communication in Malta. Since it is the Maltese sentence that is the focus of this work, sentences in English should be excluded from the corpus outright. I employed the Perl module Text-Language-Guess[13] to accomplish that. This module uses a list of stop words to identify the language of the text passed to it by producing a score which estimates (guesses) the probability of the language of said text to be that which the stop words belong to. In this case, each element in the sentence array was passed to a function which determined whether the sentence was in English and provided a language score. However, as Maltese sentences containing only a few words in English were often identified as English, an additional step was introduced to the process: for every sentence the function identified as English, the length of the sentence in words was divided by the language score to arrive at a cutoff. After some trial and error, a cutoff of 5 was found to have an acceptable precision and recall and those sentences identified as English and having a cutoff of less than 5 were removed from further processing. Such sentences were replaced by the code SNIPPED_-ENGLISH_SENTENCE.

5.3.3.6 Corpus management and querying

The data was imported into NoSketchEngine (Rychlý 2007), backend version manatee-open-2.150 and frontend version bonito-open-3.97.6, to facilitate searching and statistical processing. For this purpose, the text files were converted to one vertical file per source with one token per line, individual documents delimited with the SGML element

13 bit.ly/2iEmvgZ (last consulted on February 28th 2018)

<doc> and sentences delimited with SGML element <s>. Each <doc> element was assigned the attributes id, source, type and subtype.

The NoSketchEngine instance can be accessed at bulbul.sk/bonito2/ (login name: guest, password: Ghilm3). *BCv3* is the default corpus (under "maltiV3"); the instance also hosts the two previous versions, "maltiV1" and "maltiV2". Appendix B contains the vertical text (*.vrt) files containing all the texts in the corpus, as well as the NoSketchEngine registry files and the compiled corpus.

5.4 Enrichment

5.4.1 Part-of-speech tagging

5.4.1.1 The tagset

Table 5.7 contains the part-of-speech tagset used for the manual part-of-speech tagging of a selected subset of the *MLRSv3* and *BCv3* corpora.

5.4.1.2 Tagging decisions and their hierarchy

The tagset was compiled with the purpose of capturing the structure of Maltese as closely as possible while also reducing complexity and thus ensuring its suitability for NLP applications. The actual tag labels were chosen for readability and were inspired by those employed by the universal part-of-speech tagset (Petrov et al. 2012). The decisions made in manual tagging were largely informed by Borg and Azzopardi-Alexander 1997 and Aquilina 2007. In some cases, however, alternative analysis was applied to primarily make the tagger's (whether it's a human or a software application) job easier with consistency as the primary aim.

The following hierarchy of decisions was applied when creating the tagset:

I. Semantics
II. Morphology
III. Syntax

Criterion I lies behind such obvious choices as NOUN ("beings, things and concepts"), VERB ("words indicating motion or change of state"), ADJ ("property words") and NUM_* ("words that count"). Additionally, categories such as QUAN and FOC were created based on their semantic roles. Criterion II was used in differentiating between various types of GEN, LIL and PREP, as well as between ADJ and PART_PASS. And finally, criterion III is the primary motivator for establishing categories such as PRON_INDEF and tagging some types of words according to their role in the sentence, such as PREP vs. ADV, or maintaining a distinction between *għand* + PRON or *fi* + PRON in their roles as PREP_PRON and as VERB_PSEU. In what follows, I will describe in some detail the application of these criteria when it comes to making tagging decisions.

ID	TAG	Description	ID	TAG	Description
1	FIX_THIS	Make corrections to this token	26	NUM_WHD	number "one"
2	_IGNORE_	ignore	27	PART_ACT	active participle
3	ADJ	adjective	28	PART_PASS	passive participle
4	ADV	adverb	29	PREP	preposition
5	COMP	complementizer	30	PREP_DEF	preposition with article
6	CONJ_CORD	coordinating conjunction	31	PREP_PRON	preposition with pronoun
7	CONJ_SUB	subordinating conjunction	32	PROG	progressive particle
8	DEF	article	33	PRON_DEM	demonstrative pronoun
9	FOC	focus particle	34	PRON_DEM_DEF	demonstrative pronoun with article
10	FUT	future particle	35	PRON_INDEF	indefinite pronoun
11	GEN	genitive particle	36	PRON_INT	interrogative pronoun
12	GEN_DEF	genitive particle with article	37	PRON_PERS	personal pronoun
13	GEN_PRON	genitive particle with pronoun	38	PRON_PERS_NEG	personal pronoun with negative suffix
14	HEMM	existential verb	39	PRON_REC	reciprocal pronoun
15	INT	interjection	40	PRON_REF	reflexive pronoun
16	KIEN	the verb *kien*	41	QUAN	quantifier
17	LIL	oblique particle	42	VERB	verb
18	LIL_DEF	oblique particle with article	43	VERB_PSEU	pseudoverb
19	LIL_PRON	oblique particle with pronoun	44	X_ABV	abbreviation
20	NEG	verbal negator	45	X_BOR	unclassified
21	NOUN	noun	46	X_DIG	digits
22	NOUN_PROP	proper noun	47	X_ENG	English words
23	NUM_CRD	cardinal numeral	48	X_FOR	other foreign words
24	NUM_FRC	fractions	49	X_PUN	punctuation
25	NUM_ORD	ordinal numeral			

Tab. 5.7: Maltese part-of-speech tagset

5.4.1.3 Tags and their definition

5.4.1.3.1 _FIXTHIS_

This tag is used during manual tagging to indicate a problem (incorrect tokenization etc.) to be fixed later. It is removed once the problem is addressed.

5.4.1.3.2 _IGNORE_

This tag is used for processing document structure (<doc> and <s>) during manual tagging. It is removed for tagger training.

5.4.1.3.3 ADJ

This class includes property words which satisfy one of the following conditions:

I. They modify nouns (1);
II. form predicates in copular sentences (2);
III. and typically take feminine suffix –a, plural suffixes –i and –in or form a broken plural (3).

(1) *Il- Partit Nazzjonalista/**ADJ** qiegħed jiġġedded ...*
DEF party nationalist PROG renews itself ...
'The Nationalist Party renews itself ...'

[BCv3: illum_new.15_jannar_2015.pjklgl]

(2) *It- Tour ta' din is- sena huwa sinjifikanti/**ADJ** ...*
DEF tour GEN this.F DEF year he significant ...
'This year's Tour (de France) is significant ...'

[BCv3: ilgensillum.2011-April-12.7226]

(3) *Anita kienet lesta/**ADJ** biex toħroġ ...*
Anita she was ready-F COMP she goes out...
'Anita was ready to go out ...'

[BCv3: 2008 Loranne Vella Simon Bartolo-Wied Wirdien (Fiddien II)]

This class includes both property words that follow and property words that precede the noun. The question of what is the default position of an adjective with regard to the noun it is governed by is outside of the scope of this work, and thus the decision above was made with regard to comparatives which appear before the noun.

On morphological grounds, this class excludes PART_ACT and PART_PASS (see below). Additionally, several words are included here which do not necessarily have the semantic or morphological properties described above, but nevertheless fulfil the same role as adjectives and appear in the same position as a subclass of them, such as *aktar* "more", *istess* "same", *iżjed* "fewer", *tali* "such" and *tant* "so".

In addition to the relatively straightforward structures where ADJ modifies a NOUN or serves as the predicate or verbal complement, there are a number of others where adjectives appear on their own without a noun to modify, a subject to predicate and a verb to complement. The major types of these structures are:

i. ADJ modifies a noun that appears elsewhere in the same sentence or in a previous one (the anaphora problem). In such case, the tag is ADJ, if only for semantic reasons.

(4) *Jgħid mod u jagħmel ieħor/**ADJ***
he says way and he does other
'He says one thing and does the other'

[BCv3: netnews_lokali_20150218_jghid-mod-u-jaghmel-iehor]

ii. Adjectives which do not describe properties, but rather refer to entities having those properties and which syntactically function as nouns (i.e. ADJ>NOUN conversion), are tagged as NOUN.

(5) *Eluf ta' Maltin/**NOUN** u Għawdxin/**NOUN** żammew it-*
thousands GEN Maltese-PL and Gozitan-PL they kept DEF
tradizzjoni ...
tradition...
'Thousands of Maltese and Gozitans kept the tradition...'

[BCv3: ilgensillum.2011-Jannar-18]

iii. Superlative constructions *mill*-ADJ (ADJ in the comparative) are tagged as PREP_-DEF + ADJ:

(6) *L- Ingliżi għandhom qawl/**NOUN** mill-/**PREP_DEF** aqwa/**ADJ**...*
DEF English-PL they have saying from-DEF strongest...
'The English have the most perfect saying ...'

[BCv3: l-orizzont.96347]

iv. Adverbial constructions composed of PREP_DEF + ADJ are tagged as such, regardless of their function:

(7) *... hu u jpetpet għajnejh bil-/**PREP_DEF** goff/**ADJ**.*
... he and he blinks his eyes with-DEF awkward.
'... while he blinks awkwardly.'

[BCv3: 2012 Clare Azzopardi - Il-Każ Kważi Kollu tal-Aħwa De Molizz]

One exception to the previous rule is a group of adverbial phrases involving the adjective aħħar such as dan *il-aħħar* "lately, recently" and *fl-aħħar* "ultimately, finally" where *aħħar*, normally an adjective meaning "last", is tagged as NOUN:

(8) *Dan l- aħħar/**NOUN** m' għadnix niċċajta miegħek bħal qabel...*
this.M DEF last NEG I have-NEG I joke with you as before...
'Lately I should not be joking with you as before...'

[BCv3: 1986 Oliver Friggieri - Fil-Parlament ma Jikbrux Fjuri]

Finally, there is the issue of English words in English spelling used in Maltese sentences and fully integrated into their structure down to the morphological level (e.g. through the assimilation of the direct article). The decision has been made to tag them as if they were Maltese words (see also the entry for X_ENG below). In such cases, English nouns modifying other English nouns are also tagged as ADJ, as in (9):

(9) *B' hekk fi ftit ħin kellna ħames gas/**ADJ** turbines/**NOUN**...*
with thus in little time we had five gas turbines...
'And this way, in a short time we had five gas turbines...'

[BCv3: 20000623_324d_par]

5.4.1.3.4 ADV

Manner words which modify predicates (10) or adjectives (11):

(10) *Morna lura/**ADV** flok 'il quddiem/**ADV**!*
we went back instead of to front!
'We went back instead of forward!'

[BCv3: 20100208_190d_par]

(11) *Għalkemm il- Premier hu tqil wisq/**ADV**, l- Ewwel Diviżjoni hu aktar/**ADV** diffiċli.*
although DEF Premier he hard very, DEF first division he more difficult.
'Although the Premier (League) is very hard, the Primera división is more difficult.'

[BCv3: illum.2008-05-18.sport]

This word class is where criterion III must be applied consistently. As is immediately obvious, there is a lot of ambiguity between ADV and ADJ on one hand and ADV and PREP on the other. For example in (12), *tajjeb*, which would be normally classified as an adjective on morphological grounds alone, modifies the predicate and is thus tagged as ADV. In (13), *wara*, typically a preposition, also modifies the predicate and is therefore also tagged as ADV.

(12) *Fra Mudest beda jħossu tajjeb/**ADV** u kompla jimxi.*
brother Mudest he began he feels well and he continued he walks.
'Brother Mudest began to feel better and continued walking.'

[BCv3: 2011 Charles Casha - Mid-dinja ta' Fra Mudest]

(13) *Wara/**ADV** kien hemm riċeviment.*
after was EXIST reception.
'Afterwards, there was a reception.'

[BCv3: ilgensillum.2011-Jannar-18.3790]

Additionally, ADV can appear preceded by prepositions (such as *'il quddiem* in (10) or *s'issa* "until now"). In such case, they are also tagged as ADV.

In non-copular verbless clauses (see chapter 6, section 6.4.4.1.4) of the type *tajjeb li* "it is good that" or *żgur li* "it is certain that", it is assumed that these constructions are compositionally equivalent to structures an adverb, such as *għalhekk li* "it is so that" and *dażgur li* "it is certain that". In such cases, the tagging hierarchy applies: ambiguous word like *tajjeb* and *żgur* which could be interpreted as either ADJ or ADV are tagged as ADV (criterion III), unless they show morphological characteristics of adjectives (criterion II):

(14) *Hija ċara/**ADJ** li jibżgħu mill- konfront.*
EXPL clear-F COMP they fear from-DEF confrontation.
'It is clear they are afraid of confrontation.'

[BCv3: illum_new.9_awwissu_2016.kompla_jgerreq_ilvapur]

And finally, this category excludes focus particles/adverbs like *ukoll* and *biss* which some grammars (like Borg and Azzopardi-Alexander 1997: 83–84) classify as adverbs. Based on the analysis in Čéplö 2017, focus particles are assigned their own tag (see below).

5.4.1.3.5 COMP

This word class includes complementizers, primarily the ubiquitous *li*, but also *ma* in multiword subordinating conjunctions featuring a PREP, such as the one in (15). The etymologically related *kulma*, however, is tagged as PRON_INDEF.

(15) *Qabel ma/**COMP** nsellmu lil xulxin...*
before COMP we greet ACC each other...
'Before we greet each other...'

[BCv3: illum.2010-02-14.interview]

As Borg and Fabri (2016: 421) note, *li* "cannot be uniquely associated with complementation since it also introduces modifying (relative) clauses". However, this double syntactic role is not reflected in its part-of-speech tag and *li* is consistently tagged COMP. Same applies to subordinating conjunctions *biex* and *jekk* which in addition to this role also can serve as complementizers (Borg and Fabri 2016: 421), but are consistently

tagged CONJ_SUB (see below). The decision on which tag to use was based on considerations of frequency of both roles in *BCv3* as determined from a random sample of 100 hits where the CONJ_SUB function predominated.

5.4.1.3.6 CONJ_CORD

Coordinating conjuctions, i.e. *imma*, *inkella*, *iżda*, *jew*, *però* and *u*. Additionally, in the *kemm … kif ukoll* construction, *kemm* is also tagged as CONJ_CORD.

5.4.1.3.7 CONJ_SUB

Subordinating conjunctions, i.e. *avolja* (and its variant *allavolja*), *bħallikieku*, *bħalma*, *biex*, *billi*, *daqslikieku*, *filwaqt*, *għalkemm*, *għalli*, *għax*, *jekk*, *kieku*, *la* (as opposed to its homograph in the *la … lanqas* construction), *ladarba*, *malli*, *meta*, *milli*, *mindu*, *sabiex*, *sakemm* and *xħin*. This is another word class where the syntactic criterion III needs to be applied consistently, as there exists ambiguity between PREP and CONJ_SUB, as with *daqskemm*, *minflok* and *minħabba* which normally function as prepositions, but can assume the role of a subordinating conjunction.

There are other words that function as subordinating conjunctions, such as the aforementioned multiword expressions combining PREP and COMP (*waqt li*, *qabel ma* etc.), but those are tagged according to their constituent parts. The CONJ_SUB only includes words that connect directly without a complementizer. *filwaqt* may look like an exception since it is nearly always followed by *li*, but it can also appear on its own, hence its inclusion here.

5.4.1.3.8 DEF

This class contains the definite article *il-* in all its forms.

5.4.1.3.9 FOC

This class contains focus particles (Čéplö 2017), i.e. *anke*/*anki*, *biss*, *lanqas* (with *anqas* and *inqas* as variants where applicable), *mqar*, *saħansitra* (and its variants like *sansitra*), *ukoll*/*wkoll*. Additionally, *basta* is also tagged as FOC wherever applicable.

5.4.1.3.10 FUT

Future particles preceding verbs, i.e. *se*, *ser* and *ħa*. This also includes *għad* when it performs the same function (Vanhove 1993: 194-195):

(16) *Għada* *għad*/***FUT*** *ikollna* *bżonnu…*
tomorrow FUT we will have his need...
'Tomorrow we will need him...'

[BCv3: 1986 Oliver Friggieri - Fil-Parlament ma Jikbrux Fjuri]

The active participle of the verb *sar* in all its forms, i.e. *sejjer*, *sejra* and *sejrin*, is also tagged FUT when modifying a verb as in (17).

(17) *Imma x'* *sejjer/**FUT*** *nagħmel?*
but what FUT I will do?
'But what will I do?'

[BCv3: dilemma]

5.4.1.3.11 GEN/GEN_DEF/GEN_PRON

The genitive particle *ta'* on its own, with fused definite article and with fused pronouns (Borg and Azzopardi-Alexander 1997: 206).

5.4.1.3.12 HEMM

The existential pseudoverb *hemm* (see section 5.4.1.3.36 below), but also its less frequent alternative *hawn* (18), along with their forms with the negative suffix attached.

(18) *Għalfejn hawn/**HEMM** ħafna wirdien* *f' Malta bħalissa?*
why EXIST many cockroach-PL in Malta now?
'Why are there so many cockroaches in Malta now?'

[BCv3: 2008 Loranne Vella Simon Bartolo-Wied Wirdien (Fiddien II)]

There exists an ambiguity between ADV and HEMM for both *hemm* and *hawn*. This is resolved based on their respective syntactic function.

5.4.1.3.13 INT

Interjections such as *mela*, *le*, *iva*, *grazzi*, *prosit* and *pereżempju*, but also curses and various hard to pin down words with discourse functions, such as the particle *ja*.

5.4.1.3.14 KIEN

All the forms of the verb *kien*, including those with the negative suffix attached.

5.4.1.3.15 LIL/LIL_DEF/LIL_PRON

The oblique particle *lil* on its own, with fused definite article and with fused pronouns (Borg and Azzopardi-Alexander 1997: 195).

5.4.1.3.16 NEG

The verbal negator *ma* and its allomorph *m'*, but also *la* in the construction *la X u lanqas Y*.

5.4.1.3.17 NOUN

Words which:

I. Denote beings, objects and concepts;
II. can form plurals in *-i*, *-at*, *-iet*, *-ijiet* and *-in* or broken plurals;
III. and can be modified by the definite article or an adjective.

This word class also includes ADJ converted to nouns (5) as well as English nouns adapted to the morphology of Maltese (9).

5.4.1.3.18 NOUN_PROP

Proper nouns, i.e. names of people, places etc. Other word classes which are used as proper nouns (e.g. *It-Tlieta* as a day of the week) are also tagged as NOUN_PROP. Names of entities which are composed of generic nouns (e.g. the magazine *Il-mument* or book and movie titles) are tagged based on their constituent parts. Parts of names of Maltese localities that the tokenizer splits off (e.g. *Hal-* in *Hal-Għargħur*), titles that form parts of names (e.g. *San*, *Santu*) and syntactically analyzable parts of foreign names (e.g. *de* or *von*) are also tagged NOUN_PROP.

5.4.1.3.19 NUM_CRD

Cardinal numerals with the exception of *wieħed*. Plurals of words expressing multiples of ten (e.g. *għexieren* "tens", *mijiet* "hundreds", *eluf* "thousands" and *miljuni* "millions") and which connect to their head by means of *ta'* are tagged as NOUN.

5.4.1.3.20 NUM_FRC

Fractions, i.e. *nofs* "half", *terz* "third" and *kwart* "quarter".

5.4.1.3.21 NUM_ORD

Ordinal numerals, e.g. *ewwel* "first" and *tieni* "second".

5.4.1.3.22 NUM_WHD

The word *wieħed* "one", its feminine form *waħda* and its plural *uħud*.

5.4.1.3.23 PART_ACT

Active participles of the Arabic type (Mifsud 1995: 39) in any stem, e.g. *nieqes*, *ġej* or *imsiefer*, including their feminine forms and plural forms.

5.4.1.3.24 PART_PASS

Passive participles of both the Arabic type (Mifsud 1995: 39) and the Romance type (Mifsud 1995: 132-135).

This definition may seem trivial, but it is only so for Semitic verbs. For Romance passive participles, the morphological situation is much more complicated, since some words which are etymologically PART_PASS are indistinguishable from adjectives in their use. The membership of Romance passive participles in this word class must therefore be decided on syntactic grounds; in other words, the only true Romance PART_PASS is that which can be used in a passive clause. Consequently, both morphological and syntactic criteria are applied determining what is a proper PART_PASS and what is not. The tagging criteria therefore maintain that for a Romance candidate form to qualify as a PART_PASS, one of these conditions must be fulfilled:

I. There is a verb from which the candidate form can be derived and such a verb is in active use (however infrequent) as evidenced in *BCv3*.
II. The candidate form can be and is used (again without any regard to frequency) in the Romance passive construction with *ġie* ("the dynamic passive" in Borg and Azzopardi-Alexander 1997: 214, Vanhove 1993: 321-324).

To give an example, forms like *rikoverati* "recovered-PL" are tagged as PART_PASS because they can appear in passive constructions like *jiġu rikoverati*. Forms like *diżorganizzat* "disorganized", however, are only used attributively or predicatively the same way adjectives are. Moreover, there is no verb attested in *BCv3* from which they could be derived using any of the existing morphological processes (Mifsud 1995: 80-251). This contrasts them with forms like *interessat* "interested" or *irrabjat* "angry" which are also not used in the dynamic passive, but the verbs they are derived from (*interessa* and *rrabja*, respectively) are well attested in *BCv3*. Consequently, *diżorganizzat* is tagged as ADJ while *interessat* and *irrabjat* are tagged as PART_PASS.

There are some outlier cases, such as *separat* "separated": its source verb *ssepara* "to separate" is exceedingly rare in *BCv3* (e.g. all its imperfect forms combined come up to only 8.29 per million) and the past participle itself is predominantly used as ADJ. It is, however, also employed in phrases like *jiġi separat* "it is separated" which are rare and, more importantly, confined to legal and parliamentary documents. Since those too is a part of the Maltese language, *separati* is tagged PART_PASS as per condition ii above.

5.4.1.3.25 PREP/PREP_DEF/PREP_PRON

This class includes prepositions (Stolz and Levkovych 2017), whether on their own, with fused definite article or fused pronouns. For marginal cases which straddle the fence between ADJ/ADV and PREP like *qrib* "close" or between NOUN and PREP like *permezz* "permission", the following criterion is applied: only words that can attach di-

rectly to the noun are classified as prepositions. Consequently, *qrib* in (19) is a preposition, while *permezz* in (20) is not:

(19) *Daħal fil- kamra u resaq qrib/**PREP** Marku.*
he entered in-DEF room and he approached close Marku.
'He entered the room and approached Marku.'

[BCv3: 2010 John Bonello - It-Tielet Qamar]

(20) *Kienu mqabbdin ma' magna speċjali permezz/**NOUN** ta' xi*
they were connected-PL with machine special by means GEN some
wires.
wires.
'They were connected to a special machine by means of some wires.'

[BCv3: 2009 Carmel G. Cauchi - Il-ġrajjiet ta' Jacob Jones]

The morphological criterion can be called for support here as well, considering that like other prepositions, *qrib* can take a fused pronoun to give *qribu* "close to him, near him", while *permezz* cannot and must resort to attaching the pronoun to the genitive marker *ta'*, giving *permezz tiegħu* "by means of him/it". This is somewhat complicated by the fact that *qrib* can also behave similarly; to 137 examples of *qribu*, there are 490 examples of *qrib tiegħu* in *BCv3*. Nevertheless, *permezz* never takes fused pronouns and so the distinction stands.

The postposition *ilu* is also tagged PREP.

5.4.1.3.26 PROG

This word class contains the verbal particles *qed* and *qiegħed* (in all its forms) which denote the progressive nature of an action or process (Vanhove 1993: 113-129). As with FUT, only those occurrences of *qiegħed* that modify verbs are tagged as PROG, those that have different functions (either a copula or an existential predicate, see Chapter 6, sections 6.4.4.1.3 and 6.4.4.1.5) are tagged as PART_ACT.

5.4.1.3.27 PRON_DEF/PRON_DEM_DEF

Demonstrative pronouns *dan*, *din*, *dawn* and *dawk*, including their forms with fused definite article (*dal-*, *daċ-* etc.).

5.4.1.3.28 PRON_INDEF

Indefinite pronouns, i.e. words that express absolute quantification ("all" or "nothing") or unspecified quantification ("some") and are in complementary distribution with nouns or (typically adverbial) noun phrases. These include *għalxejn* "for nothing", *ħadd*

"someone / (in negative sentences) no one", *kollox* "everything", *kulma* "everything that", *kulħadd* "everyone", *kulmin* "everyone who" and *xejn* "nothing", including *ilkoll* "everything" which also does double duty as a determiner and *ħaddieħor* "someone else".

This category contains negative words that trigger *x*-dropping like *ħadd* and *xejn* (see Lucas 2014) and for this reason, I also included here words which behave the same way, but normally do not have a quantifying meaning and so cannot serve the same syntactic functions as the other members of this word class. The primary example of such words is *mkien* "place" which behaves like *ħadd* in that in negative sentences, it means "nowhere" and thus can only be an adverbial, never a subject or an object.

5.4.1.3.29 PRON_INT

Interrogative pronouns (Borg and Azzopardi-Alexander 1997: 210-212), e.g. *x'*/*xi* (as opposed to its quantifier homophone, see below), *kif*, *min*, *fejn*, *kemm*, *kemm-il*, *liema*, *xiex* etc. This class also includes the special form of interrogative third person personal pronouns *inhu*, *inhi* and *inhuma*. In this case, criterion III was not applied and so no distinction is made between PRON_INT as interrogatives and PRON_INT as complementizers.

5.4.1.3.30 PRON_PERS

Independent personal pronouns (Borg and Azzopardi-Alexander 1997: 195), whether long or short (Stolz and Saade 2016), including those with the interrogative suffix *-x* attached.

5.4.1.3.31 PRON_PERS_NEG

Independent personal pronouns, whether long or short, with the negative suffix *–x*. This does not include their forms with its homophone/homograph interrogative suffix, since the distinction here is quite clear: PRON_PERS_NEG either contain the negator *ma* as a prefix (e.g. *mhux*, *mhix*, *mhumiex*), or they are preceded by it (typically in its reduced form *m'*, e.g. *m'iniex* or *m'intix*). Interrogative personal pronouns are not accompanied by the negator and are therefore tagged as PRON_PERS.

5.4.1.3.32 PRON_REC

The reciprocal pronoun *xulxin* "each other".

5.4.1.3.33 PRON_REF

The reflexive pronouns *nnifs-* and *ruħ-* in all their forms, plus the word *stess*.

5.4.1.3.34 QUAN

5.4.1.3.35 VERB

Verbs which exhibit full forms in both the prefixal conjugation (i.e. the Semitic imperfect) and the suffixal conjugation (the Semitic perfect) (cf. Mifsud 1995 and Spagnol 2011).

5.4.1.3.36 VERB_PSEU

This class contains pseudoverbs, i.e. those words that function as verb-like predicates, but do not take either imperfect or perfect affixes (Peterson 2009). These fall into two groups: group 1 contains pseudoverbs that obligatorily take what are for all intents and purposes conjugation suffixes which are identical to attached pronouns/clitics. This group contains *beħsieb-* "to intent to", *donn-* "to appear as if", *fi-* "to contain", *għad-* "to continue to", *kell-/għand-/ikoll-* "to have", *għodd-* "to almost X", *il-* "to have been X", *jisem-* "to be named", *qis-* "to be like" and *waħd-* "alone", including *fiħsieb-*, a low-frequency variant of *beħsieb-*, missing from Peterson's list. Group 2 then contains those pseudoverbs that only take the negative suffix *-x*, i.e. *hemm* "there is" (and its variant *hawn*, missing from Peterson's list) and *għad* "to be still" (also missing from Peterson's list). Group 2 also contains *tantx*, the negated form of the adverb *tant* "so much", not on Peterson's list and only included here on morphological grounds.

Both groups of pseudoverbs are assigned the tag VERB_PSEU, with two exceptions: *waħd-* is tagged as ADV since this is its actual function; clauses where it acts as the predicate are syntactically equivalent to copular sentences with an adverb as the predicate. Additionally, *hemm/hawn* are assigned their own tag HEMM due to their special status as predicates in existential sentences (see Chapter 6, section 6.4.4.1.5).

In this context, there is an ambiguity between PREP_PRON and VERB_PSEU for *fi-* and *għand-*, between FUT and VERB_PSEU for *għad* and between PREP and VERB_PSEU for one form of *il-*, *ilu*. This is resolved based on their respective syntactic functions.

5.4.1.3.37 X_ABV

Abbreviations (*Dr.*, *PN*), acronyms (*NATO*) and contractions (*Sta* in *Sta Venera* for *Santa Venera*).

5.4.1.3.38 X_BOR

Anything that is not anything else, like nonsense characters, OCR artefacts, Roman numerals etc. This also includes single letters used e.g. as list delimiters, section titles and in mathematical expressions.

5.4.1.3.39 X_DIG

Digits (0-9 and beyond). This includes digits 11-19 with the suffix *-il* attached, as well as numbers with decimals (e.g. 29.38).

5.4.1.3.40 X_ENG

This tag is used for English words not assimilated into the morphological and syntactic structure of Maltese. In practice, the following rule was applied: English noun phrases analyzable as a combination of NOUN and ADJ are tagged as such (21). The actual tag is selected based on the function of the word in the sentence, or rather on the function of Maltese words (whether Semitic or Romance) which appear in the same position.

(21) *Dawn huma d- double/**ADJ** standards/**NOUN** tal- Kap tal- Oppożizzjoni, saħaq Dr Muscat.*
these.M they DEF double standards GEN-DEF leader GEN-DEF opposition, insisted Dr. Muscat.
'These are the double standards of the leader of the opposition, insisted Dr. Muscat.'

[BCv3: illum_new.3_april_2016.liskema_ta_dejn_sigriet]

If, however, a particular sequence of tokens displays English syntax, say by containing a preposition or a verb with its arguments, they are all tagged X_ENG (22):

(22) *Ma kienx worth/**X_ENG** it/**X_ENG**.*
NEG he was-NEG worth it.
'It wasn't worth it.'

[BCv3: inewsmalta-lul.14.2013.1840-8692]

5.4.1.3.41 X_FOR

This tag is used for words from languages other than Maltese or English, as with the German terms below:

(23) *In- nom Verarschung/**X_FOR** ġej mill- verb Verarschen/**X_FOR**...*
DEF noun Verarschung come.PART.ACT from-DEF verb Veraschen...
'The noun *Verarschung* comes from the verb *verarschen*...'

[BCv3: maltarightnow.2012-7-5.58-9983862]

5.4.1.3.42 X_PUN

Punctuation, i.e. all non-alphabetical non-numeral characters that are not a part of a token (i.e. hyphens and apostrophes).

5.4.1.4 Manual tagging

This version of the tagset is the product of the merger of the tagset used in *bulbulistan malti v1* (2013) and the tagset used in *MLRS Korpus Malti v2.0* (2012). For both corpora, a set of texts was manually annotated with parts of speech and used to train an automated tagger. In the merger, I reviewed both sets of texts, made changes to the manual annotation wherever necessary and added a number of new texts, expanding the entire manually tagged corpus to 111 texts comprising 109,053 tokens.

5.4.1.5 Automated tagging

The manually annotated files were used to test a number of automated part-of-speech taggers. SVMTool v 1.3.2 (Giménez and Márquez 2004) was determined to provide the best accuracy, in addition to several other advantages like advanced configurability and the ability to include dictionaries to improve tagging. Testing has shown that for Maltese and the tagset described here, the SVMTool tagger provides the best performance with the following settings:

```
W = 5 2 #window definition (size, core position), default 5 2
F = 1 100000 #feature filtering, default 2 100000
X = 11 #unknown words, default 3
Dratio = 0.001 #default 0.001
...
do M0 LR CK:0.0635 CU:0.14
```

With a 90:10 random train:test split and 10 test runs, the average accuracy of the tagger is 97.35% (see Table 5.8), i.e. state of the art (Jurafsky and Martin 2009: 189).[14]

The full dataset was used to train a model with the same settings which was then used to automatically tag all the texts in *BCv3*.

14 The same assessment is made in the 3rd edition of the classic hanbook still in progress in late 2017, see web.stanford.edu/ jurafsky/slp3/10.pdf, last consulted on February 28th 2018.

Run	Known	Ambiguous known	Unknown	Overall
1	99.0003%	92.6063%	80.7799%	97.3345%
2	98.7738%	90.4950%	84.1842%	97.4903%
3	98.8009%	91.2169%	82.9596%	97.4421%
4	98.7440%	90.9548%	81.3084%	97.2387%
5	98.9339%	91.5191%	82.3584%	97.4292%
6	98.8164%	90.9382%	79.9193%	97.0536%
7	98.8944%	91.7505%	80.8000%	97.2418%
8	98.8458%	91.5477%	82.1503%	97.3576%
9	99.0249%	92.8728%	81.0011%	97.4733%
10	99.0588%	93.0818%	82.0641%	97.4845%
Average	98.89%	91.70%	81.75%	97.35%

Tab. 5.8: SVMTool part-of-speech tagging accuracy

6 Maltese Universal Dependencies Treebank v1

6.1 Introduction

This chapter describes a Maltese treebank based on the Universal Dependencies treebank annotation standard (UD; Nivre, de Marneffe et al. 2016a), henceforth referred to as MUDTv1. The version designation v1 refers to both the status of MUDT, as well as to the fact that the MUDT annotation scheme is an extension of Universal Dependencies v1 (UD v1; Nivre, Ginter et al. 2014; Nivre, de Marneffe et al. 2016) as opposed to its most recent version 2.1 (UD v2; Nivre, Ginter et al. 2016; Nivre, Agić et al. 2017).

6.2 Universal Dependencies

6.2.1 Why Universal Dependencies?

The Universal Dependencies annotation has emerged as the de facto standard in syntactic annotation for NLP purposes. This is evidenced not only by its adoption as such by the industry (Andor et al. 2016), but also by the fact that it serves as the entry point into advanced NLP for languages where little to no such resources exist; this, in turn, can be shown by the growth of the UD treebank database from 10 languages in January 2015 to 60[1] in November 2017 (UD v2.1 release; Nivre, Agić et al. 2017) with further 6 planned for 2018. It is my belief that existing standards (at least those involving all things digital) should be adhered to, no matter how flawed they are, and this belief alone would have been reason enough for me to choose UD as the basis for the syntactic annotation of a Maltese treebank. In this case, however, I have done so gladly and without any reservations: UD is a remarkably well organized and implemented project with goals and aims that are nearly identical to mine – it uses traditional linguistic labels with notions behind them extensible as needed, it lends itself to fairly rapid and reasonably consistent human annotation and UD-annotated treebanks have been shown to better enable high-accuracy automated parsing than some existing formats (Antomonov 2015). To use it to compile a treebank for my own goals and by doing so allow others to include Maltese in other tasks for which UD-annotated corpora were intended (primarily cross-linguistic comparison) therefore makes perfect sense. And if this justification isn't sufficient, I have another one: I have compiled and annotated this treebank on my own time, with my own resources. So there.

1 This count includes a Maltese treebank. MUDTv2 will be included in the November 2018 release of UD v2.2.

6.2.2 Levels of annotation and record format

UD is the child of two initiatives, the Stanford dependencies (de Marneffe, MacCartney and Manning 2006; de Marneffe and Manning 2008; de Marneffe, Dozat et al. 2014) and the universal part-of-speech tagset (Petrov, Das and McDonald 2012); it is these two types of information that form the basis of the UD annotation (McDonald et al. 2013). In UD v1 (Nivre, Ginter et al. 2016), morphology information was added to the UD standard based on the morphological layer in the HamletDT treebank (Zeman 2008) bringing the total of annotation levels in UD v1 to 10. These are encoded in the CoNLL-U format where treebanks are stored in plain text UTF-8-encoded files with one word per line (word line), empty lines marking sentence boundaries and hashtags (#) marking comments. Each word line consists of 10 tab-separated fields which contain the actual annotation. Table 6.1 (adopted with minor modifications from Nivre, Ginter et al. 2014) describes the use for each field.

Field	Layer	Description
1	ID	Word index, integer starting at 1 for each new sentence; may be a range for tokens with multiple words.
2	FORM	Word form or punctuation symbol.
3	LEMMA	Lemma or stem of word form.
4	UPOSTAG	Universal part-of-speech tag drawn from the revised version of the Google universal POS tags.
5	XPOSTAG	Language-specific part-of-speech tag.
6	FEATS	List of morphological features from the universal feature inventory or from a defined language-specific extension.
7	HEAD	Head of the current token, which is either a value of ID or zero (0).
8	DEPREL	Universal Dependency relation to the HEAD (root iff HEAD = 0) or a defined language-specific subtype of one.
9	DEPS	List of secondary dependencies (HEAD-DEPREL pairs).
10	MISC	Any other annotation.

Tab. 6.1: Levels of annotation in UD / CoNLL-U format

For a treebank to qualify as a UD treebank, only fields 1, 2, 4, 7 and 8 are required (Nivre, Ginter et al. 2014; Nivre, Ginter et al. 2016). The remaining ones can be left unspecified which is marked by an underscore ("_") as the CoNLL-U format does not allow empty fields. In what follows, I will discuss the details of UD annotation simultaneously with its application to Maltese.

6.3 Maltese UD annotation

6.3.1 ID

A sequential integer ID. The UD guidelines define words as not phonological or orthographic, but as syntactic units ("it is important to note that the basic units of annotation are syntactic words (not phonological or orthographic words)"; Nivre, Ginter et al. 2014). The UD specification thus allows for ranges to be used to mark multiword tokens, such as clitics attached to verbs. While this is applicable to Maltese and would in fact be useful, I have not implemented it in MUDTv1 due to difficulties with the morphological analysis of Maltese encliticized verbs (see also Chapter 5, section 5.3.3.4).

6.3.2 FORM

The token as defined in Chapter 5, section 5.3.3.4.

6.3.3 LEMMA

This is not used in MUDTv1.

6.3.4 UPOSTAG: Universal part-of-speech tags

The UD v1 annotation scheme extended Petrov's original tagset of 12 coarse part-of-speech tags (Petrov, Das and McDonald 2012) to 16 which are listed in Table 6.2 below.

Open class words	Closed class words	Other
ADJ	ADP	PUNCT
ADV	AUX	SYM
INTJ	CCONJ	X
NOUN	DET	
PROPN	NUM	
VERB	PART	
	PRON	
	SCONJ	

Tab. 6.2: Universal part-of-speech tags

In MUTDv1, the UPOS tags are not used and their position is occupied by the Maltese-specific part-of-speech tags. A mapping between the two tagsets is relatively

straightforward with the exception of those part-of-speech tags that combine two word classes, like PREP_DEF and LIL_DEF.

6.3.5 XPOSTAG: Maltese-specific part-of-speech tags

See Chapter 5, section 5.4.1, for a full list and description of Maltese part-of-speech tags.

6.3.6 FEATS: Maltese morphological features

6.3.6.1 General

In MUDTv1, this field is left empty. Nevertheless, some preliminary work on morphological annotation has been done in advance of MUDTv2 which is described in this section. In line with the status of MUDTv1 as an expansion of UD v1, the following discussion will refer to UD v1 morphological features, highlighting the relevant changes in UD v2 whenever appropriate.

The morphological data in UD treebanks is encoded in the form of pairs of features (i.e. morphological properties) and the values these features can take. This is recorded in the $\mathrm{Feature1}{=}\mathrm{Value1}|\mathrm{Feature2}{=}\mathrm{Value2}|\ldots$ format where the pipe character (|) separates individual feature/value pairs which are ordered alphabetically by feature name. Table 6.3 contains the full list of morphological features in UD v1.

Lexical features	Inflectional features	
	*Nominal**	*Verbal**
PronType	Gender	VerbForm
NumType	Animacy	Mood
Poss	Number	Tense
Reflex	Case	~~Aspect~~
	Definite	Voice
	Degree	Person
		Negative

Tab. 6.3: UD v1 morphological features

In Maltese, the lexical features are already encoded in the Maltese-specific part-of-speech tags for specific subclasses of pronouns and numerals since these are, with one exception, the only word classes where they appear (this also applies to UD v2). The only exception to this are (nouns with) possessive suffixes which are addressed below.

As for inflectional features, only aspect is not applicable to Maltese, since it is marked by a complex system of auxiliary verbs and particles (Vanhove 1993: 39-274). It is therefore not used for Maltese and marked with strikethrough in the table above.

In what follows, the use of the remaining features and the values they can take in Maltese is described in detail.

6.3.6.2 Lexical features: Poss

The lexical feature Poss can only take a single value, Yes and is only applicable to tokens tagged GEN_PRON.

6.3.6.3 Nominal feature: Gender

The Gender feature can take one of two values, Gender=Masc and Gender=Fem. This feature is applied to nouns (NOUN, NOUN_PROP) including symbol classes that can stand for nouns (X_ABV), adjectives (ADJ), numerals including symbol classes that stand for those numerals (NUM_FRC, NUM_ORD, NUM_WHD, X_DIG), demonstrative pronouns (PRON_DEM, PRON_DEM_DEF), personal pronouns, including those attached to other word classes (GEN_PRON, LIL_PRON, PREP_PRON PRON_PERS, PRON_PERS_-NEG, PRON_REF), and verbs and verb-like word classes (KIEN, VERB, VERB_PSEU, PART_ACT, PART_PASS).

For nouns, adjectives and demonstrative pronouns, Gender is mandatory in all forms. For personal pronouns, GEN_PRON, LIL_PRON, VERB, VERB_PSEU and the KIEN, the distinction is only made in the third person singular.

6.3.6.4 Nominal feature: Animacy

The Animacy feature is only applied to word classes NOUN, NOUN_PROP and X_ABV (when the latter refers to entities normally expressed by the former two) and can take one of the three values: Animacy=Anim, Animacy=Nhum and Animacy=Inan. The first category refers to human beings (real or fictional) or to things and concepts when personified, while the third is used for objects, institutions, concepts and alike. The second category is used for those objects, institutions and concepts that are inanimate, but treated as animate. The stereotypical examples for this category in Maltese would be *Malta* and *Partit Nazzjonalista* (*PN*) which refer to the island (and the country) and a political party, respectively, yet they both take the direct object marker *lil* normally reserved for persons "and other expressions high on the scale of animacy" (Borg and Azzopardi-Alexander 1997: 137). This behavior, i.e. differential object marking, is the sole motivator behind establishing Animacy as a feature for Maltese.

6.3.6.5 Nominal feature: Number

The Number feature can take one of two values, Number=Sing and Number=Plur, and is applied to nouns (NOUN, NOUN_PROP) including symbol classes that can stand for nouns (X_ABV), adjectives (ADJ), the numeral *wieħed* (NUM_WHD), demonstrative pronouns (PRON_DEM, PRON_DEM_DEF), personal pronouns, including those attached to other word classes (GEN_PRON, LIL_PRON, PREP_PRON PRON_PERS, PRON_PERS_NEG,

PRON_REF), and verbs and verb-like word classes (KIEN, VERB, VERB_PSEU, PART_ACT and PART_PASS).

6.3.6.6 Nominal feature: Case

The Case feature is applied to GEN, GEN_DEF, GEN_PRON, LIL, LIL_DEF and LIL_PRON and can take one of three values, Case=Gen (which is automatically assigned to GEN, GEN_DEF and GEN_PRON), Case=Dat and Case=Acc. The latter two are used with LIL, LIL_DEF and LIL_PRON only, depending on their exact syntactic function.

6.3.6.7 Nominal feature: Definite

The Definite feature is applicable to the markers of definiteness DEF, GEN_DEF, LIL_-DEF, PREP_DEF and PRON_DEM_DEF where it automatically takes the value Definite=Def. Additionally, this feature is applied to NOUN, but only those that are a part of the so-called construct state (Borg and Azzopardi-Alexander 1997: 75-76) as the *nomen regens*. This NOUN is definite by virtue of being in a construct state, but cannot bear any of the markers of definiteness listed above (Borg and Azzopardi-Alexander 1997: 112). In such cases, the NOUN token is marked with Definite=Red.

6.3.6.8 Nominal feature: Degree

The Degree feature is applicable to ADJ and ADV only and can take one of two values: Degree=Pos for adjectives and adverbs in the basic form (i.e. non-comparative) and Degree=Cmp for comparative adjectives and adverbs. The value Degree=Sup (superlative) is not used in Maltese since it is morphologically identical to the comparative and the only distinction is in the definiteness which is encoded as a feature in a separate word (Borg and Azzopardi-Alexander 1997: 75).

6.3.6.9 Verbal feature: VerbForm

The VerbForm feature is applied to KIEN, PART_ACT, PART_PASS, VERB and VERB_PSEU and can take one of two values: VerbForm=Fin (which is automatically assigned to all forms of KIEN, VERB and VERB_PSEU) and VerbForm=Part (which is automatically assigned to all forms of PART_ACT and PART_PASS).

6.3.6.10 Verbal feature: Mood

The Mood feature to KIEN, VERB and VERB_PSEU only and can take one of two values, Mood=Ind Mood=Imp. Simply put, imperative forms of KIEN and VERB are assigned feature Mood=Imp, all the remaining ones, including all VERB_PSEU, are marked as Mood=Ind.

6.3.6.11 Verbal feature: Tense

The Tense feature applies to KIEN, VERB and VERB_PSEU and can (with one exception) take one of two values, Tense=Pres and Tense=Past. Tense=Pres is used for the prefixal conjugation (the imperfect) forms of KIEN and VERB, as well as for all VERB_PSEU with the exception of *kell-*. Tense=Past is applied to the suffixal conjugation (the perfect) of KIEN and VERB_PSEU, as well as the VERB_PSEU *kell-*.

The only exception to this is the VERB_PSEU *ikoll-* which is the only verb in Maltese that refers to the future in and of itself (Borg and Azzopardi-Alexander 1997: 367). As such, it takes the value Tense=Fut by default.

6.3.6.12 Verbal feature: Voice

The Voice feature is applied only to VERB, VERB_PSEU, PART_ACT and PART_PASS. It can take one of two values, Voice=Act (which is automatically assigned to PART_ACT and VERB_PSEU) and Voice=Pass (automatically assigned to PART_PASS). For Verb, Voice=Pass is applied to verbs in one of the passive derived stems (Spagnol 2011: 109, Borg and Azzopardi-Alexander 1997: 213) and Voice=Act is assigned to all the others.

6.3.6.13 Verbal feature: Person

The Person feature is applied to personal pronouns, including those attached to other word classes (GEN_PRON, LIL_PRON, PREP_PRON PRON_PERS and PRON_PERS_NEG) and to finite verbs and verb-like words (KIEN, VERB and VERB_PSEU). This feature can take on values Person=1, Person=2 and Person=3.

6.3.6.14 Verbal feature: Negative

The Negative feature can take on values Negative=Pos and Negative=Neg and is applied to word classes HEMM, KIEN, PRON_PERS_NEG, VERB and VERB_PSEU. Simply put, the presence of the negative suffix is automatically encoded as Negative=Neg and its absence as Negative=Pos. The PRON_PERS_NEG class stands out here: theoretically, the PRON_PERS word class could also have this feature encoded (with the value Negative=Pos), however, I decided against it, since the sole function of this morphological feature is to indicate negation on a predicate or a copula. In addition to being the latter, words in the PRON_PERS class can also function as subjects of clauses and those cannot be subject to negation.

In UD v2, this feature is replaced by Polarity. All of the above applies without change.

6.3.6.15 Note: Clitics

In MUDTv2, both verbal clitics and suffixed possessive pronouns will be split off as separate tokens. Verbal clitics will be marked with Case, Gender and Number; suffixed

possessive pronouns will be marked with the Poss lexical feature and Gender and Number.

6.3.6.16 Morphological features: summary

Table 6.4 below summarizes the features which specific tags can take:

Word class	Gender	Animacy	Number	Case	Definite	Degree	VerbForm	Mood	Tense	Voice	Person	Negative
ADJ	×		×			×						
ADV						×						
DEF					×							
GEN				×								
GEN_DEF				×	×							
GEN_PRON	×		×	×							×	
HEMM												×
KIEN	×		×				×	×	×	×	×	×
LIL				×								
LIL_DEF				×	×							
LIL_PRON	×		×	×							×	
NOUN	×	×	×		×							
NOUN_PROP	×	×	×									
NUM_FRC												
NUM_ORD	×											
NUM_WHD	×		×									
PART_ACT	×		×				×					
PART_PASS	×		×				×					
PREP_DEF					×							
PREP_PRON	×		×								×	
PRON_DEM	×		×									
PRON_DEM_DEF	×		×		×							
PRON_PERS	×		×								×	
PRON_PERS_NEG	×		×								×	×
PRON_REF	×		×									
VERB	×		×				×	×	×	×	×	×
VERB_PSEU	×		×				×	×	×	×	×	×
X_ABV	×	×	×									
X_DIG	×											

Tab. 6.4: UD v1 morphological features in Maltese

6.3.7 HEAD: Head of the current word

This field contains the word ID of the head (governor) of the current word or 0 if the word is the sentence root.

6.3.8 DEPREL: Maltese universal dependency relations

This field contains the UD relation label. Section 6.4 below describes in detail the application of these labels to the syntactic structure of Maltese.

6.3.9 DEPS: Enhanced dependency graph

This is not used in MUDTv1.

6.3.10 MISC: Any other annotation

This is not used in MUDTv1.

6.4 Maltese UD relations, or: a sketch of Maltese syntax

6.4.1 Introduction

In this section, I describe the adaptation of the UD v1 annotation scheme to Maltese and the decisions I made while applying the UD v1 relation labels to the structure of Maltese sentence. Although this process may seem easy and uncomplicated, it actually amounts to compiling a rough description of Maltese syntax. A large number of the annotation decisions I made was informed by previous works on Maltese syntax, especially Borg and Azzopardi-Alexander 1997 and Vanhove 1993. In others, however, I applied my own analysis of the syntactic structures in question, one whose primary aim was not necessarily full descriptive adequacy, but rather compatibility with the UD v1 standard, ease of annotation, simplicity (or parsimony) and most of all, consistency. Experience with annotation has taught me that it is better to be consistently wrong than inconsistently right; the former is much easier to correct. Additionally, a full description of the grammar of Maltese is still a desideratum and consequently, some of the phenomena I deal with have received little to no attention in scholarly literature on Maltese. In such cases, I based my analysis on the way equivalent phenomena were handled in other UD v1 treebanks (primarily Hebrew, for obvious reasons), as well on my investigation of the Maltese phenomena in *BCv3*. I will use this space to lay out the decisions I made in such cases and the reasoning behind them in detail.

6.4.2 General principles of syntactic annotation in UD v1

The primary goal of UD – maximum parallelism between typologically diverse languages – is achieved by consistently annotating the same syntactic relations in the same way. A set of general principles has been compiled to that end. The following list is a summary of these principles based on Nivre, Ginter et al. 2014:

I. UD relations form a tree with a single root
II. No empty / zero nodes or relations
III. UD relations are relations between content words
IV. UD relations are as flat as possible
V. Basic UD relations can be extended to account for the particularities of any language

Principle I is one of the fundamentals of dependency linguistics (see chapter 1, section 1.3.2.2 and references therein). Principles II and III are intertwined – if the basis of the syntactic is the content word (principle III), then there is no need to postulate empty nodes as some dependency grammars which treat function words as heads have to (cf. Oxborne and Maxwell 2015: 248). As for principle III, Nivre, de Marneffe et al. (2016: 1662-1663) also note that this is consistent with the fundamentals of dependency grammar as outlined by Tesnière: Tesnière differentiates between two types of words, "full words" which are "charged with semantic function" and "empty words" which are "simple grammatical tools" (*Elements*, Chapter 28, §1-3). Dependency relations hold between nuclei (*Elements*, Chapter 22, §11) and only full words can form the core of a nucleus (*Elements*, Chapter 31, §8). More importantly, however, this principle is a prerequisite for the primary purpose of UD, i.e. cross-linguistic comparison, since different languages often express the same relationship – say, a locative noun phrase modifying a verb – in different ways: a preposition for the likes of English and German, a suffix for Finnish and Slovak and no marker at all for Chinese and Maltese. Principle IV is ultimately nothing but a simple trick designed to simplify the annotator's work – in fact, the UD annotation guidelines (Nivre, Ginter et al. 2014 and Nivre, Ginter et al. 2016) frame it more as a recommendation than a hard rule. And finally, principle V is an obvious acknowledgement of the vast differences between languages; in practical terms, it is typically implemented by establishing subcategories for existing relations.

6.4.3 Rules of syntactic annotation in UD v1

The rules of syntactic annotation in UD v1 (Nivre, Ginter et al. 2014; Nivre, Ginter et al. 2016) can be summarized as follows:

I. We distinguish four types of dependents: nominals, clauses, modifiers and function words

   a. Multi-word expressions are a special form of modification.
II. We differentiate core arguments from non-core arguments.
   a. We differentiate subjects of passives from other subjects.
III. We differentiate predicate dependents from nominal dependents.
IV. We differentiate clauses which inherit a subject from a higher clause from clauses with their own internal subject.

Rule I establishes the fundamental classification of dependencies which is ultimately based on the structure of catenae and the valency of their heads: nominals have nouns or pronouns at their heads while clauses have verbs or other types of predicates as their heads; both can be modified by all types of dependents and both thus can be recursive. Modifiers and function words, on the other hand, cannot be recursive and only allow a very limited and specific set of dependents.

Rule II seeks to address – or rather avoid – the old argument vs. adjunct issue. The UD v1 guidelines (Nivre, Ginter et al. 2014) describe the UD taxonomy as "centered around the fairly clear distinction between core arguments (subjects, objects, clausal complements) versus other dependents" and leave it at that. While the general idea of core and non-core dependents is a sound one, the devil – as is his want – is in the details, especially those of core predicate arguments. The original concept of core arguments vs. non-core (oblique) dependents is derived from the Lexical Functional Grammar (LFG, de Marneffe and Manning 2008: 4586 and Dalrymple 2011: 8-27). The criteria LFG uses for the distinction are somewhat vague, to a large extent language-specific (Dalrymple 2001: 23) and in any case wholly irrelevant since the analysis here is not based on LFG. To resolve the issue and adequately and consistently make "the clear distinction" between core arguments and non-core dependents, I have established one additional rule for annotation:

V. Classification of predicate arguments is based on the valency frame of the predicate.

I will discuss the detailed application of this rule in the section on core arguments below. For now, suffice it to say that this is the reason I have also amended the list of UD v1 dependencies with categories I considered necessary for an adequate and consistent description of predicate dependents.

Rule IIa is established to adequately represent sentences where the subject is not the agent, i.e. the stereotypical passive sentences.

Rule III is a corollary of Principle I: if a distinction is made between nominals and clauses (because of the special status of the verb), a distinction must be made between clauses which modify nominals and clauses which modify other clauses.

And finally, Rule IV is another one inherited from the Stanford Dependencies where the distinction between two types of complement clauses is ultimately traceable to the concepts of XCOMP and COMP in the Lexical Functional Grammar (Dalrymple 2001: 24-

26). And while it is not necessary for any of the stated purposes of UD v1, it is useful for some aspects of typological and syntactic analysis, including the one conducted here, especially when it comes to the problem of auxiliaries (on which see below).

Table 6.5 below summarizes the UD v1 dependency relations, ordered by type using the format of UD v2 and including special types of dependents and relations. This list has been adapted to Maltese by adding relations to account for the specific properties of Maltese and for the type of analysis I employ here; they are marked with an underline. Dependencies marked with an asterisk (*) perform double duty as both modifiers of nominals and dependents of the predicate; they therefore appear in both respective rows.

	Nominals	Clauses	Modifier words	Function Words
Core arguments	nsubj	csubj		
	nsubjpass	csubjpass		
	dobj	ccomp		
	iobj	xcomp		
	nmod:obj			
Non-core dependents	~~nmod~~*	advcl	advmod*	aux
	nmod:agent		discourse	auxpass
	nmod:advmod			cop
	vocative			mark
	expl			neg
	dislocated			part
Nominal dependents	nmod*	acl	amod	det
	nmod:poss		advmod*	case
	appos			case:det
	nummod			
Coordination	MWE	Loose	Special	Other
conj	compound	list	foreign	punct
cc	mwe	parataxis	goeswith	root
	name		remnant	dep
			reparandum	

Tab. 6.5: UD v1 relations adapted to and extended for Maltese

In what follows, I will discuss each of those and their application to Maltese, using a top-down order starting with root and illustrating them with examples from *MUDTv1* and (whenever more convenient due to clause length) *BCv3* using dependency graphs.[2]

2 A note on formatting: in these graphs, punctuation tokens are joined to their nearest non-punctuation token to the left. This is in contrary to the principles of UD and actual annotation in MUDTv1 and is done for reasons of space only. To consult the examples cited from MUTDv1, see section 6.5.4.

6.4.4 Maltese UD relations

6.4.4.1 root

6.4.4.1.1 General remarks

This section discusses predicates, i.e. roots of clauses in general; everything said here applies not only to the label root which stands for the root of the sentence (the main clause), but also to all types of clauses which are dependents of the sentence root, any other clause or any other word. Their respective roots can be labeled csubj, csubjpass, advcl, acl, xcomp, ccomp, parataxis or conj. In what follows, all these (including clauses with their root labeled root) will be termed UD clauses types.

In Maltese, clauses can also be divided into five types based on which word class (i.e. Maltese-specific part-of-speech tag) they have as their root and the resulting structure of the clause:

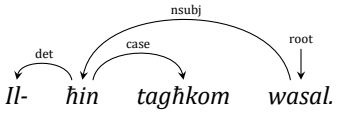
I. Verbal clauses
II. Copular clauses
III. Non-copular verbless clauses
IV. Existential clauses
V. Non-expletive subjectless clauses

Each of them and their particulars will be discussed below. This discussion is relevant not only for the general descriptive purposes and UD v1 annotation of syntactic relations, but also for the analysis that is the primary aim of this work.

6.4.4.1.2 Verbal clauses

Verbal clauses are clauses which have a verb or a pseudoverb as the root. The definition of a verb and a pseudoverb for the purposes of MUTDv1 is the same as the one applied for part-of-speech tagging, i.e. the primary criterion is morphological. Verbal clauses are thus those that have as their root a token tagged as VERB (1), VERB_PSEU, PART_ACT (with the exception of *qiegħed* in all its forms, see section 6.4.4.1.3 on copular clauses below) and PART_PASS.

(1) *Il- ħin tagħkom wasal.*
DEF time your.PL he came.
'Your time has come.'



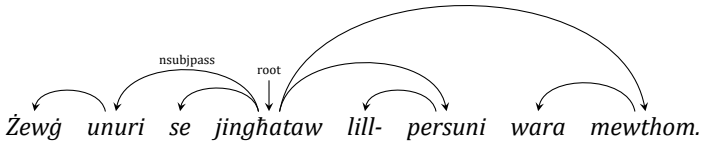
[MUDTv1: 49b_04F09]

Based on the morphological properties of these word classes and the effects on the structure of the clause they have, verbal clauses can be further divided into two groups:

i. Active clauses
ii. Passive clauses

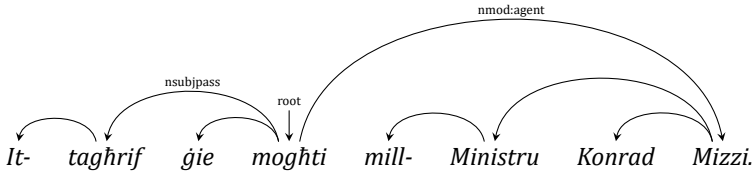
In line with the verb-central nature of the syntactic analysis here, passive clauses in MUDTv1 are primarily defined by their root: a clause is passive, if the root is either a VERB with the morphological feature Voice set to Pass (i.e. the verb is in one of the passive stems, cf. Spagnol 2011: 109 and Borg and Azzopardi-Alexander 1997: 213) as in (2) or the root is a PART_PASS (3):

(2) *Żewġ unuri se jingħataw lill- persuni wara mewthom.*
two honor-PL FUT they are given DAT-DEF person-PL after their death.
'Two honors will be given to people after their death.'



[BCv3: inewsmalta-dic.13.2014.1005-22170]

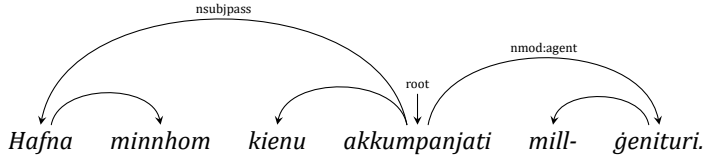
(3) *It- tagħrif ġie mogħti mill- Ministru Konrad Mizzi.*
DEF information it came given from-DEF minister Konrad Mizzi.
'The information was given by Minister Konrad Mizzi.'



[BCv3: inewsmalta-ott.29.2013.1257-11045]

This analysis is problematic in the second part, as in Maltese, there are two types of analytic passive constructions containing a passive participle: the so-called "dynamic passive" (Borg and Azzopardi-Alexander 1997: 214, see also Vanhove 1993: 321-324) which combine PART_PASS with the passive auxiliary *ġie*, and the so-called "stative passive" (Borg and Azzopardi-Alexander 1997: 214, Vanhove 1993: 318-320) where the place of *ġie* is taken by KIEN:

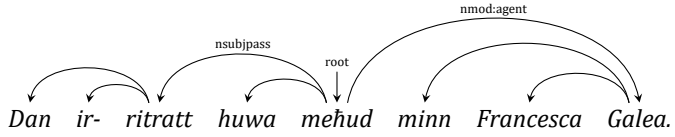
(4) *Hafna minnhom kienu akkumpanjati mill- ġenituri.*
many from them they were accompanied-PL from-DEF parent-PL
'Many of them were accompanied by parents.'



[BCv3: 2008 Loranne Vella Simon Bartolo-Wied Wirdien (Fiddien II)]

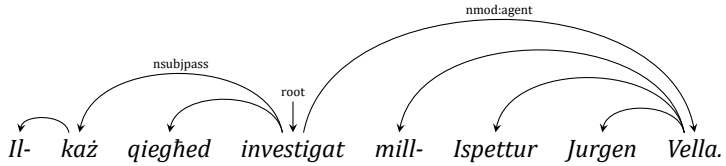
This is the only type of the stative passive that has so far been described in literature on Maltese. Considering the basic function of KIEN as a copula and various other types of Maltese copular clauses (see section 6.4.4.1.3 below), it is hardly surprising that there are equivalent structures where the place of KIEN is taken by PRON_PERS (5), *qiegħed* and its forms or even left empty (7):

(5) *Dan ir- ritratt huwa meħud minn Francesca Galea.*
this DEF picture he taken from Francesca Galea
'This picture was taken by Francesca Galea'



[BCv3: arblu]

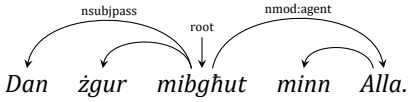
(6) *Il- każ qiegħed investigat mill- Ispettur Jurgen Vella.*
DEF case COP investigated from-DEF inspector Jurgen Vella.
'The case is being investigated by inspector Jurgen Vella.'



[BCv3: inewsmalta-jan.24.2014.0806-12927,newspaper]

(7) *Dan żgur mibgħut minn Alla.*
this.M certainly sent from God.

'This is surely sent by God.'



[BCv3: 2012 Claire Azzopardi - Kidane]

What all the types of the stative passive have in common is a problem with their status as passive clauses: PRON_PERS, *qiegħed* and KIEN also serve as copulas and PART_PASS can also be used in the same function as an adjective, both atributively and predicatively. As a result, an ambiguity arises between passive clauses composed of a passive subject with no copula/PRON_PERS/*qiegħed*/KIEN + PART_PASS on one hand and copular clauses with PART_PASS as the adjectival predicate with no copula/PRON_PERS/*qiegħed*/KIEN on the other. This ambiguity can sometimes be resolved on syntactic grounds alone, as with PART_PASS like *interessat* "interested" or *irrabjat* "angry": both are proper passive participles (as per Chapter 5, section 5.4.1.3.24, condition I) since there exist verbs they are derived from (*interessa* and *rrabja*, respectively); yet they are not attested in the dynamic passive construction in *BCv3*. As such, the clauses they feature in with PRON_PERS, *qiegħed* or KIEN as direct dependents are considered copular clauses, not passive clauses. In other cases, however, the ambiguity can only be resolved semantically: as both an overt (lexical) subject and an agent noun phrase are optional in such clauses, their mere absence is not indicative of anything. One must thus analyze the semantics of the verb or rather its valency frame and this is where Rule V comes into play: to arrive at a decision on whether a clause with a PART_PASS as a root is passive or not, one must determine whether the particular PART_PASS can feature an agent noun phrase. Here as in the criteria for determining what is a PART_PASS (Chapter 5, section 5.4.1.3.24), no regard is paid to frequency and so for example of the 2291 clauses combining KIEN with *magħluq* "to be closed" in *BCv3*, only 13 feature an agent noun phrase, but this is enough to analyze such clauses as passive. Verb valency is also why ultimately clauses featuring PART_PASS such as *irrabjat* or *interessat* cannot be passive: as evident from *BCv3*, *irrabja* is monovalent, *interessa* is reflexive.

This discussion is relevant to the purposes of the this work for two reasons: first, to provide the background for the analysis of copular clauses which present a number of difficulties in a language like Maltese (see below). Secondly and more importantly, Rule IIa of UD v1 annotation requires that subjects of passives be analyzed differently from other subjects. The definition of a passive clause above which relies on the possibility of the clause featuring an overt agent noun phrase is established with this in mind.

To summarize, passive clauses are those that have as their root

a. a VERB with the morphological feature Voice set to Pass,
b. a PART_PASS with *ġie* as a direct dependent, or

c. a PART_PASS with a PRON_PERS, one of the forms of *qiegħed*, a KIEN and/or a nsubjpass as a direct dependent, but only if the valency of the verb the PART_PASS is derived from allows an agent noun phrase.

All other verbal clauses, including those with a VERB_PSEU as their root, are treated as active.

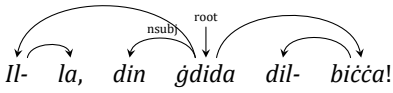
6.4.4.1.3 Copular clauses

In UD v1, copular clauses are treated differently from verbal clauses, in that the copula is not considered the head of the clause, but rather a dependent of the lexical predicate, be it a noun, adjective, adverb, a prepositional phrase or anything else (Nivre, Ginter et al. 2014). This analysis was adopted to account for languages which do not require an overt copula like Russian or Hungarian. Maltese is one of those languages, at least when it comes to the present tense where a copula is not obligatory (Borg and Azzopardi-Alexander 1997: 49) and only sentences with a past timeframe require the use of KIEN (Borg and Azzopardi-Alexander 1997: 52). Consequently, there are four types of copular sentences in Maltese (Borg and Azzopardi-Alexander 1997: 53):

i. No copula
ii. Personal pronoun as the copula
iii. Present participle *qiegħed* as the copula
iv. KIEN as the copula

Type (i) is a typical example of what traditional grammars of Semitic languages (e.g. Muraoka 2005: 82-86 for Syriac or Zewi 1994 for Biblical Hebrew) refer to as the nominal sentence. In Maltese, copula-less copular clauses like (8) are used for identity, attribution and location (cf. the classification of copular clauses in Dixon 2010: 159):

(8) *Il- la, din ġdida dil- biċċa!*
INT INT, this.F new-F this.F-DEF bit
'Whoa, this is new, this thing!'

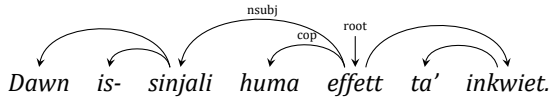


[BCv3: 2009 Loranne Vella Simon Bartolo - Il-Ġnien tad-Dmugħ (Fiddien III)]

Maltese type (ii) copular clauses are somewhat unique in the North African linguistic milieu, but not entirely unprecendented: similar constructions can be found in Syrian Arabic (Berlinches 2016: 138), Anatolian Arabic (Lahdo 2009: 172-173) and especially Cypriot Maronite Arabic, another variety of Arabic heavily influenced by an Indo-

European language (Borg 1985: 135). (9) is a standard Maltese example showing its use in identity (or equative, cf. Borg and Spagnol 2015) constructions:

(9) *Dawn is- sinjali huma effett ta' inkwiet.*
these.M DET signal-PL they effect GEN unrest
'These signals are the result of unrest.'



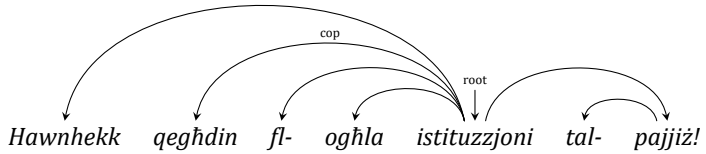
[BCv3: 2001 Peter Caruana - Għarfien il-Pjanti]

Type (iii) copular clauses, contrary to the previous type, do have counterparts in dialects of Arabic related to Maltese (Benkato and Pereira 2015). In Maltese, they are typically used for the expression of:

a. "locative predications", (Borg and Azzopardi-Alexander 1997: 49-50, 53; Borg and Spagnol 2015)
b. "a (temporary) role" (Borg and Azzopardi-Alexander 1997: 50, 53, 143)
c. "a temporary state" (Borg and Azzopardi-Alexander 1997: 51, 53)

Example (10) below illustrates the first of these, the locative usage:

(10) *Hawnhekk qegħdin fl- ogħla istituzzjoni tal- pajjiż!*
here COP-PL in-DEF highest institution GEN-DEF country!
'Here we are in the highest instution in the country!'

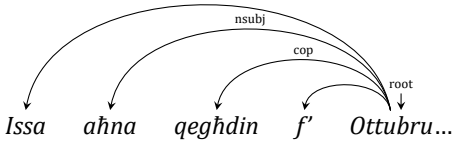


[BCv3: 20131001_056d_par]

It goes without saying that this usage can be extended to metaphorical locations, such as state of mind (e.g. *qiegħed f'estasi "*to be in an ecstasy"), a relation (*qiegħed f'kollegament telefoniku "*to be in a telefonic connection") or a situation (*qiegħed f'qagħda prekarja "*to be in a precarious situation"). It is perhaps from the last named usage that a previously under-described subtype of such sentences has emerged, one where the predicate is not locative, but temporal:

(11) *Issa aħna qegħdin f' Ottubru...*
now we COP-PL in October...

'Now we're in October...'



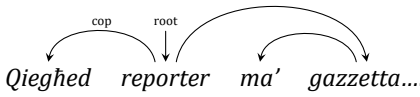
[BCv3: ilgensillum.2012-Frar-5.10878]

Such use of *qiegħed* seems to be limited to a handful of predicates, like names of calendar months, divisions of the year (weeks, months, seasons), unspecified periods of time (e.g, *fażi"*phase") and adverbials of time (Borg and Spagnol 2015). Borg and Azzopardi-Alexander (1997) do not describe this subtype of *qiegħed* clauses; they do, however, describe the other chief usage of such clauses as copular clauses of identity and attribution involving temporary (or transient) roles and properties. Example (12) taken from Borg and Azzopardi-Alexander (1997: 50) illustrates the former (original glossing and translation retained):

(12) *Pietru qiegħed l-eżaminatur*
Peter located - 3sg.m the-examiner
'Peter is temporarily the examiner.'

This type of predication is sometimes referred to as stage-level predication and contrasted with individual-level predication, where the former denotes temporary properties or states with "an inherent end, that is, telicity" (Olsen 2014: 48), whereas the latter describes "inherent qualities" (Roby 2009: 39). Borg and Azzopardi-Alexander (1997) argue for this interpretation of type (iii) copular clauses implicitly, whereas Camilleri and Sadler (2018) do so explicitly, describing *qiegħed* as a stage-level copula and noting its parallels in other varieties of Arabic. However, this interpretation of copular clauses like (12) is contradicted by two types of such constructions involving identity and attribution. The first one is best exemplified by (13):

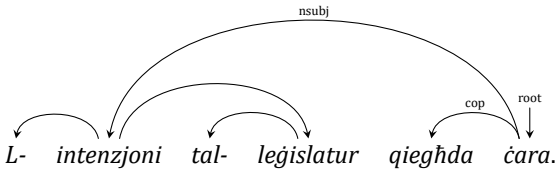
(13) *Qiegħed reporter ma' gazzetta...*
COP reporter with newspaper...
'He is a reporter with a newspaper...'



[BCv3: 2011 Trevor Żahra - Qamar Aħdar]

These constructions (which are by no means frequent) typically feature an occupational designation (e.g. *impjegat* "employee", *ministru* "minister", *għassa* "guardian" and *gowl* "goalkeeper") or a noun denoting membership (e.g. *membru* "member" and *parti* "part") as the predicate, invariably non-definite, thus indicating that these are truly clauses of identity (Borg and Azzopardi-Alexander 1997: 49-50). And while some of these could be conceivably considered transient or non-permanent in temporary sense (e.g. government posts such as minister are occupied for a limited time only), this is hardly the case with occupational designations. An alternative explanation is required, for which consider copular clauses of attribution featuring *qiegħed* such as (14):

(14) *L- intenzjoni tal- leġislatur qiegħda ċara.*
DEF intention GEN-DEF legislator COP-F clear-F
'The legislator's intention is clear.'



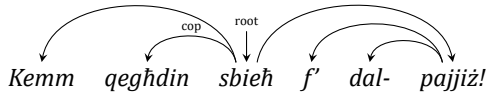
[BCv3: 20040426_015d_kun]

Here once again, there can be no talk of temporary or transient nature of the legislator's intention: the discussion involves a legal measure which captures the intent of the legislator in a rather permanent manner. What is at question is its interpretation at a particular point in time, i.e. at the point of speaking. And this is in my view the correct analysis of the semantics of these constructions: they express durative identity or attribution as perceived at the point of the utterance. In (13) and similar clauses, this is a person's occupation or role at the time of speaking; in (14), this is a state of things as perceived as the time of speaking. This analysis for (13) is very well compatible with all types of occupational designations and memberships, typically denoted by NOUN predicates, as well as with locative constructions denoted by ADV, prepositional phrases and NOUN. The analysis for (14) is also compatible with other ADJ predicates denoting non-permanent qualities and states, such as *korrett* "correct", *preżenti* "present", *komdu* "comfortable" or *kwiet* "quiet" (as in Borg and Azzopardi-Alexander 1997: 51).

This conclusion is somewhat complicated by two factors: first, the imprecise definition of non-permanent qualities. For example, Borg and Spagnol (2015) describe *qiegħed* as not ocurring with "permanent states" like *intelliġenti* "intelligent", only with "transient states, spanning over time" like *kuntent* "satisfied" and *marid* "ill".Data from *BCv3*, however, contradicts Borg and Spagnol's conclusion: there are type (iii) copular clauses with predicates like *sbieħ* "pretty.PL" (15) which is surely a quality on par with

intelligence when it comes to permanency – unless, of course, one wishes to argue that beauty does have an inherent end associated with age, which is hardly relevant in (15).

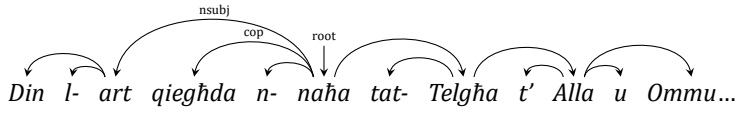
(15) *Kemm qegħdin sbieħ f' dal- pajjiż!*
how COP-PL pretty.PL in this.M-DEF country!
'How pretty we are in this country!'



[BCv3: it-torca.10383]

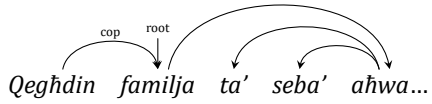
The second factor complicating the conclusion above involves copular clauses featuring *qiegħed* which contain predicates denoting types of location (16) or identity (17) that are the very opposite of transient:

(16) *Din l- art qiegħda n- naħa tat- Telgħa t' Alla u Ommu...*
this.F DEF land COP DEF side GEN-DEF hill GEN God and his mother...
'This (plot of) land is on the side of Telgħa t' Alla u Ommu...'



[BCv3: 20050712_290d_par]

(17) *Qegħdin familja ta' seba' aħwa...*
COP-PL family GEN seven brother.PL...
'We are a family of seven brothers...'



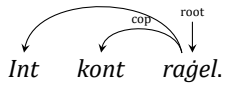
[BCv3: illum.2008-05-04.sport]

In previous analyses, copular constructions such as (16) were subsumed under the locative use of type (iii) clauses: both Borg and Azzopardi-Alexander (1997: 163, 166) as well as Borg and Spagnol (2015) cite a few examples, but do not make any comments regarding their nature. Constructions like (17), on the other hand, have so far gone unnoticed. A thorough analysis of this subtype of copular clauses is well beyond the scope of this work, so I will leave the matter with a hypothesis on the semantic restriction: for

type (iii) copular clauses denoting permanent location, the subject is invariably a part of the landscape such as a piece of land or a building. Type (iii) copular clauses denoting permanent relationships are an extention of such clauses describing occupations or memberships, the only difference being that this time, the membership is intrinsic. Whether this hypothesis is correct remains to be seen.

And finally, type (iv) copular clauses are the closest equivalent Maltese has to the Standard Average European copula. Such clauses are used for any type of copular constructions (cf. Dixon 2010: 159-179); example (18) illustrates a copular clause of identity.

(18) *Int kont raġel.*
you you were man
'You were a man.'



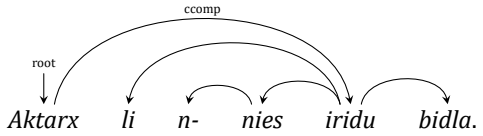
[BCv3: 2010 Immanuel Mifsud - Fl-Isem tal-Missier (U tal-Iben)]

This brief overview of copular clauses is provided here to inform the following discussion of other clause types and the assignment of various syntactic relations and will be referred to whenever appropriate.

6.4.4.1.4 Non-copular verbless clauses

There is a clause type where the root is not a verb or a pseudoverb, but which do not feature a copula or an actual subject; instead, they invariably govern a complement clause (Borg and Azzopardi-Alexander 1997: 31). In the most conspicuous subtype of such clauses, the root is a single word:

(19) *Aktarx li n- nies iridu bidla.*
probably COMP DEF people they want change.
'It is probable that people want a change.'

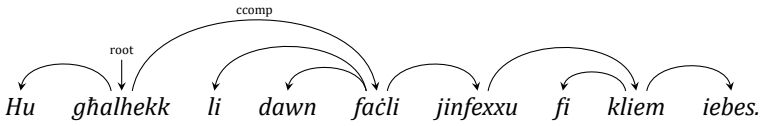


[BCv3: illum.2007-09-30.toniupeppi]

In MUTD v1, *aktarx* is thus labeled as root while the complement clause is tagged as ccomp since it has its own subject distinct from that of the main cause (see section 6.4.4.4.4).

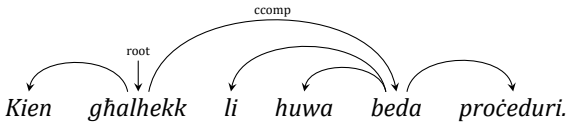
In this type of clauses, the root is typically an ADV of attitude like *ċertament* "certainly", *dażgur* "definitely" or *għalhekk* "thus, for this reason". There is, however, a subtype in which the clause also contains a PRON_PERS (20) or KIEN (21) as a dependent of the ADV:

(20) *Hu għalhekk li dawn faċli jinfexxu fi kliem iebes.*
he thus COMP those.M easy-PL they are vented in words ugly.
'It is thus so that they are easily vented in ugly words.'



[BCv3: ilgensillum.2011-Awwissu-10.8954]

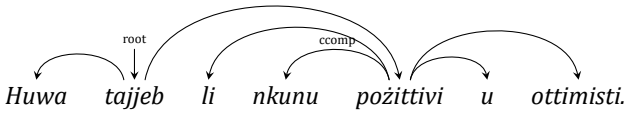
(21) *Kien għalhekk li huwa beda proċeduri.*
he was thus COMP he he began proceedings.
'It was for this reason that he began the proceedings.'



[BCv3: qorti-maltarightnow.com.2015_03_13_the-times-jinghata-ragun-dwar]

The situation is further complicated by three more subtypes of such clauses. The third subtype is one where the word class membership of root of the main clause is somewhat ambiguous, as with *tajjeb* below:

(22) *Huwa tajjeb li nkunu pożittivi u ottimisti.*
he good/well COMP we are positive-PL and optimist-PL
'It is good that we are positive and optimists.'



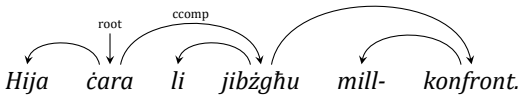
[BCv3: l-orizzont.52022]

As shown in chapter 5, section 5.4.1.3.4, *tajjeb* can modify both nouns (as adjectives do) and predicates (as adverbs do), so in this case, it is difficult to be sure what it is: on the one hand, it could be seen as taking the same syntactic role as *għalhekk* in the

examples above and thus fall into the ADV word class; on the other hand, it could also be analyzed as ADJ with the personal pronoun agreeing with it in gender.

In my analysis for the purposes of part-of-speech tagging, I adopted the former view; however, there is a fourth subtype of these clauses where it is absolutely obvious by the morphology of the root that the root is an adjective (23):

(23) *Hija ċara li jibżgħu mill- konfront.*
she clear-F COMP they fear from-DEF confrontation.
'It is clear they are afraid of confrontation.'



[BCv3: illum_new.9_awwissu_2016.kompla_jgerreq_ilvapur]

And the rabbit whole goes even deeper, as there is a fifth subtype of such clauses which can feature a noun (24) or even a prepositional phrase (25) as the predicate:

(24) *Hi ħasra li l- MUT ħadet dan il- pass.*
she pity COMP DEF MUT she took this.M DEF step.
'It's a pity that the MUT took this step.'

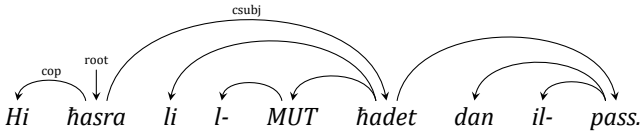
[BCv3: l-orizzont.39635]

(25) *Kien fl- imsemmi xahar li darba minnhom qabadha 'synus'...*
it was in-DEF mentioned month COMP time from them he grabbed her sinus infection...
'It was in the aforementioned month that she came down with a sinus infection...'

[MUDTv1: 05_05J01]

Examples like (24) and (25) throw two kinds of doubt on the analysis above: first, the agreements between *hi* and *ħasra* in (24) serves – along with the agreement between *hi* and *ċara* in (23) – as further proof that such clauses could be analyzed as copular. Secondly, the semantics of both (24) and (25) suggest the alternative of analyzing their respective structures as not two clauses, but a single one featuring a clausal subject (csubj, see section 6.4.4.4.1 below) with a non-verbal predicate and a copula, e.g. for (26):

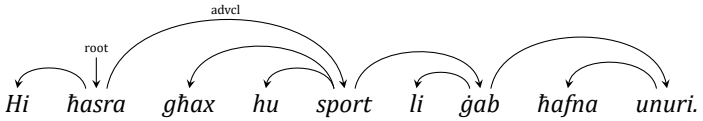
(26) *Hi ħasra li l- MUT ħadet dan il- pass.*
she pity COMP DEF MUT she took this.M DEF step.
'It's a pity that the MUT took this step.'



[BCv3: l-orizzont.39635]

However, for at least a subset of these clauses (those featuring a single noun as the root), the csubj analysis is contradicted by examples like (27) where the adverbial clause makes it clear that *Hi ħasra* is a clause of its own.

(27) *Hi ħasra għax hu sport li ġab ħafna unuri.*
she pity because he sport COMP he brought much honor-PL.
'It's a pity because this is a sport that has brought many honors.'



[BCv3: illum.2009-09-20.sport]

This leaves us, by and large, with the first option, to analyze these clauses as copular. There are three arguments against this, all involving the purported subjects of such clauses. The first one is syntactic: as examples like (19) show, the subject is not necessary in such clauses. This may be unproblematic if they were just type (iv) copular clauses and examples like (21), since the subject is encoded within the copular verb and so an overt subject (whether a pronoun or a noun) is not necessary anyway. But what are we to do with examples like (23) or (24)? In main copular clauses with PRON_PERS as the only dependent of the predicate, the PRON_PERS is invariably interpreted as the subject, not the copula; and if the subject is removed, it is questionable whether they are copular clauses at all (cf. Borg and Azzopardi-Alexander 1997: 139-140).

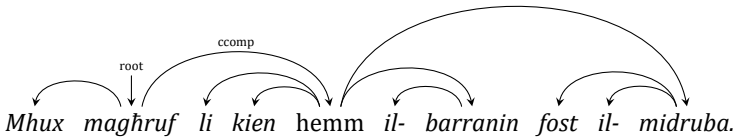
Secondly, even if these clauses were indeed copular, there are problems with their semantic analysis: those that feature an adjective (23) or a noun (27) as the predicate would be copular clauses of attribution or identity, those featuring a prepositional phrase like (25) would be locative, both standard types of copular clauses (Dixon 2010: 159). What, however, is one to make of a purportedly copular clause featuring a non-locative adverb as (20) does?

And this brings us to the semantic analysis of the predicate dependent in such clauses: while PRON_PERS and KIEN may appear to assume the place of a subject, they

do not satisfy its semantic role. There is a simple test to determine this: for pronouns, there is no antecedent; for KIEN, there is no way to add an overt subject.

The conclusion to reach here is that these are impersonal clauses with an optional expletive subject in the present timeframe and a mandatory expletive KIEN in the past or the future timeframe. Fabri is right in noting that "Maltese lacks expletive pronouns" (Fabri 2010: 794), but that does not stop Maltese from repurposing 3rd person pronouns for that function. The same is then true of their negated counterparts as the one in (28) – which, nota bene, features a PART_PASS as the root – and even of KIEN: as pronoun a and a verb inflected for person, both contain a subject and thus can also fulfill the role of an expletive subject.

(28) *Mhux magħruf li kien hemm il- barranin fost il- midruba.*
NEG known COMP he was EXIST DEF foreigner-PL among DEF wounded-PL.
'It is not known whether there were foreigners among the wounded.'



[BCv3: netnews_internazzjonali_20140715_jitilfu-hajjithom]

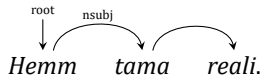
Note that this type of clauses is distinct from clauses with dummy subjects (Borg and Azzopardi-Alexander 1997: 53-54): while those also satisfy the overt subject test cited above, their root is unmistakably a verb and a such, they fall under the umbrella of verbal clauses.

Borg and Azzopardi-Alexander briefly note the existence of non-copular verbless clauses under the heading of "subjectless sentences", but describe them as only featuring "masculine singular form, if they have a normal inflection" (Borg and Azzopardi-Alexander 1997: 60). In this somewhat lengthy aside, I have been able to refine Borg and Azzopardi-Alexander's analysis and highlight the variation in the structure and syntax of such clauses not only to expand a previously under-described construction, but also to inform the discussion of one type of non-core dependents in UD v1, expl (see section 6.4.4.5.3 below).

6.4.4.1.5 Existential clauses

In terms of syntactic structure, existential clauses are clauses in which the position of the root is occupied one of the tokens tagged as HEMM, i.e. *hemm*, *hawn* or their negated counterparts *hemmx* and *hawnx* (see Peterson 2009 and chapter 5, section 5.4.1.3.12):

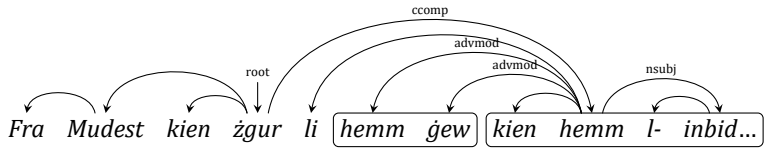
(29) *Hemm tama reali.*
EXIST hope real.
'There is real hope.'



[BCv3: illum_new.23_dicembru_2014.tidied_ilprobabbilt]

At first sight, this type of clause could be analyzed as a locative copular clause with the adverbs of place *hemm* "there" or *hawn* "here" as the root and no copula. A closer look at their semantic and morphological properties reveals that this analysis doesn't hold. First, such clauses do not actually express any locative relationship; this is evident from examples like (29) above which have subjects that are not located in space (or time), as well as examples like (30) below where a location that needs to be expressed must be provided separately (and the example was chosen precisely because it features the same word in a different syntactic function to highlight the distinction).

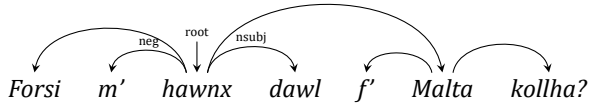
(30) *Fra Mudest kien żgur li hemm ġew kien hemm l- inbid...*
brother Mudest he was certain COMP there inside AUX EXIST DEF wine...
'Brother Mudest was certain that there was wine in there...'



[BCv3: 2012 Charles Casha - Fra Mudest]

The second argument for establishing this type of clause as separate from copular clauses involves negation: negation in copular clauses is expressed either using PRON_-PERS_NEG (types i, ii and iii, see section section 6.4.4.8.5) or by negating the copula KIEN. Contrary to that, HEMM bears its own negation:

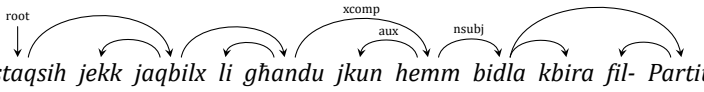
(31) *Forsi m' hawnx dawl f' Malta kollha?*
maybe NEG EXIST-NEG light in Malta all-her
'Maybe there is no light in all of Malta?'



[BCv3: 2012 John A Bonello - Is-Sitt Aħwa]

And finally, there is a third syntactic argument for treating clauses with HEMM at the root as different from copular clauses: in copular clauses, the subject agrees with the copula and/or the predicate in gender and number whenever possible. This is not the case in clauses like (32) below:

(32) *Nistaqsih jekk jaqbilx li għandu jkun hemm bidla*
I ask-ACC.3SGM if he agrees-INTR COMP he has he is EXIST change
kbira fil- Partit.
big-F in-DEF party.
'I ask him if he agrees that there should be a big change in the Party.'



[MUDTv1: 22_02J03]

The subject is clearly feminine, and yet the KIEN dependent of the clause root (as well as its governor clause root *għandu*) are masculine. In contrast, all clauses featuring the respective feminine form *tkun* + *hemm* are copular clauses where the latter word is an ADV.

These three properties alone are enough to set these types of clauses apart from copular clauses. The analysis of their semantics and use then makes it clear that they are existential clauses, i.e. clauses expressing propositions "about the existence or the presence of someone or something" (cf. McNally 2011: 1830-1831).

The etymology and current split use of *hemm* as a locative adverb and as an existential predicate establish Maltese as one of the languages where the dedicated existential predicate ("proform" in Bentley 2015a: 2) is originally locative, most notably its Tunisian relative (where *ṯamma* behaves almost identically, Ritt-Benmimoun 2014: 114) and its Romance neighbors (Bentley 2015b: 103-106). The question of how and when this happened is a fascinating one, but will need to be left for another time and place. It is, however, to some extent relevant for the present purpose, as there still seem to exist borderline cases, like (33):

(33) *Il-CAPAC hija hemm biex tagħti appoġġ.*
it CAPAC EXIST in order to she gives support.
'The CAPAC is there to provide support.'

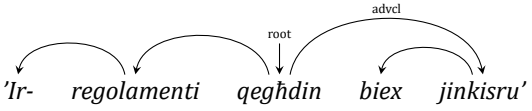
[MUDTv1: 30_01P05]

This example is one of 508 instances of such constructions (3rd person PRON_PERS + *hemm/hawn*) in *BCv3* and while it is very clearly existential, the remaining 507 clauses in *BCv3* are either clear locatives or allow both locative and existential interpretation.

A detailed analysis of such clauses is well beyond the scope of the task at hand and so in light of the low frequency and nature of these constructions, I decided to annotate them as copular clauses, with the intent to revisit that decision in MUDTv2.

And finally, there is one more type of existential clause, one featuring any of the forms of the active participle of *qagħad* as the root which typically (but not always) governs an adverbial clause of purpose introduced by *biex* "in order to" (34):

(34) *'Ir- regolamenti qegħdin biex jinkisru'*
'DEF rule-PL EXIST-PL in order to they are broken'
'The rules are there to be broken.'



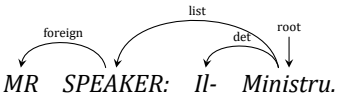
[BCv3: ilgensillum.2012-Lulju-23.16115]

Here once again one could attempt to analyze such clauses as copular, but their structure, i.e. the absence of an actual copular predicate, makes it obvious they are not, and even a cursory analysis of their use makes it clear that they too are existential clauses. This role of *qiegħed/qiegħda/qegħdin* is another one of under-described uses of the present participle of *qagħad* discussed in section 6.4.4.1.3 above. As such, it is worthy of a closer look and a more detailed analysis, e.g. as to whether there is a distinction in semantics between HEMM and *qiegħed* existential predicates (existence vs. presence). For now, suffice it to say that *qiegħed* existential clauses are not represented in MUDTv1. Should they be encountered in the future development of MUTDv1, PART_ACT *qiegħed/qiegħda/qegħdin* will be marked as root as in the example above, based on the analogy with sentences featuring a copula governing a ccomp (see section 6.4.4.4.4 below); the adverbial clause would of course be annotated as such.

6.4.4.1.6 Non-expletive subjectless clauses

This clause type encompasses all clauses that do not fall into any of the preceding types; more specifically, their root is not a VERB, VERB_PSEU or HEMM and at the same time, it does not have or cannot have an actual or an expletive subject as a dependent (hence the name). These include single-token clauses, fragments, clauses consisting of noun phrases with (35) or without dependents and so on.

(35) *MR SPEAKER: Il- Ministru.*
Mr. speaker: DEF minister.
'Mr. Speaker: The Minister.'



[MUDTv1: 38_02P06]

In some cases, especially those involving an ADJ, an ADV or a prepositional phrase as the root, the clauses can also be analyzed as copular, depending on the context (Borg and Azzopardi-Alexander 1997: 139-140). For example, the ADJ that forms the root of (36) clearly refers to one of the two participants in the conversation in which this sentence is uttered and it thus can (contrary to the definition above) take the 2nd person singular pronoun *int* as a subject.

(36) *Lesta?*
ready-F?
'Ready?'

[MUDTv1: 49b_04F09]

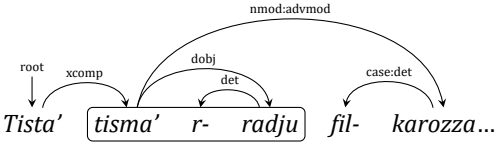
For the purposes of the analysis herein, only those clauses with a root other than VERB, VERB_PSEU and HEMM that have a subject or a copula will be considered copular; examples like (36) will be not.

6.4.4.2 Core arguments: Valency frame

6.4.4.2.1 Introduction

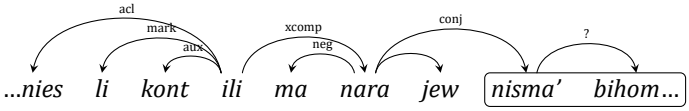
It became obvious very early in the annotation process that the "fairly clear distinction between core arguments ... and other dependents" (UD v1 annotation Rule II, see section 6.4.3 above) is anything but, and that in addition to the general guidelines provided in the UD documentation (whether UD v1 in Nivre, Ginter et al. 2014 or UD v2 in Nivre, Ginter et al. 2016), Maltese-specific definitions these relations must be provided. This is especially true of "direct object" where an urgent need emerged to come up with a practical definition that goes beyond the general categories of "the second most core argument of a verb" or "proto-patient". For an example, consider the arguments of the verb *sema'* "to hear, to listen" pair of structures like (37) and (38).

(37) *Tista' tisma' r- radju fil- karozza...*
you can you listen DEF radio in-DEF car...
'You can listen to the radio in the car...'



[MUDTv1: 21_06J02]

(38) *...nies li kont ili ma nara jew nisma' bihom...*
...people COMP I was I was for some time NEG I see or I hear with them...
'... people whom I hadn't seen or heard about...'



[MUDTv1: 21_06J02]

The noun marked only for definiteness in (37) is a relatively straightforward example of what is traditionally referred to as the direct object (e.g. Aquilina 1959: 341). What, however, is one to do with the prepositional phrase argument in (38)? The preposition *bi* typically introduces instruments, but this cannot be the interpretation here; modulo some semantic variation (nota bene the English translation), these two clauses describe the same physical process, the reception of sound waves in a person's ear, and its metaphorical extensions. And since they are semantically equivalent, should the two verbal dependents that denote the source of those waves also be considered equivalent in terms of their syntactic relationship to the verb? And what about other such structures like *kellem* "to speak" + NOUN / *tkellem* + *ma'* + NOUN, *af* "to know" + NOUN / *af* + *b'* + NOUN or *iltaqa'* "to meet" + NOUN / *ltaqa'* + *ma'* + NOUN, not to mention verbs like *nduna* "to notice" which only take nominal arguments introduced by *bi*?

This sort of variation in verbal arguments has not gone unnoticed by scholars of Maltese, most notably by Borg and Azzopardi-Alexander, who remark on it in an analysis of verbal arguments in Maltese (1997: 54-57). Their analysis operates with the traditional concepts of subject, direct object and indirect object. For directs objects, they note that they "receive a case marker (preceding them) morphologically related to the directional preposition *lil* depending on their position in the animacy hierarchy" (Borg and Azzopardi-Alexander 1997: 55). For the indirect object, section 1.2.1.2.3 (Borg and Azzopardi-Alexander 1997: 55) notes that "[an] argument can be made for maintaining a separate category 'indirect object'", but ultimately reaching the following conclusion (Borg and Azzopardi Alexander 1997: 55, emphasis in the original):

> In general, an indirect object is optional since three-place verbs like **ta** 'he gave', **bagħat** 'he sent', **kiteb** 'he wrote' can also be used with a direct object only.

In the next section, 1.2.1.2.4-5 on the types and combinations of verbal arguments, they go on to point out the behavior of some verbs which "allow an argument with the preposition **bi**, which can hardly be interpreted as 'instrument'" (Borg and Azzopardi-Alexander 1997: 57, emphasis in the original). As an example, they give the verb *ħa* "to take" which "combines with the **bi** argument, subject and direct object in [one example],

and subject and indirect object in [another example]" (Borg and Azzopardi-Alexander 1997: 57, emphasis in the original).

This, however, is as far as Borg and Azzopardi-Alexander take their discussion of the issue. At this point, the UD v1 annotation guidelines offered a clear way out: *radju* in (37) would be annotated as dobj while the prepositional phrase in (38) would be labeled nmod. I rejected this solution for various reasons, chief among them the existence of verbs like *nduna* mentioned above: one of the purposes of this work is to investigate the order of the verb and the object and as Borg and Azzopardi-Alexander note, there is no reason not to consider the arguments of *nduna* (and *sema'* and other verbs) introduced by the preposition *bi* objects. To implement this, a consistent and reliable way of identifying verbal dependents that count as objects needed to be found. The solution is already hinted at by Borg and Azzopardi-Alexander in the discussion above when they speak of "two place verbs" and "three-place verbs" (Borg and Azzopardi-Alexander 1997: 55) and it is a concept that has played a crucial role in dependency grammar from the very beginning (*Elements*, Chapters 97 through 119): valency.

Tesnière's original definition of valency is as follows:

> The verb may therefore be compared to a **sort of atom**, susceptible to attracting a greater or lesser number of actants, according to the number of bonds the verb has available to keep them as dependents. The number of bonds a verb has constitutes what we call the verb's **valency** (*Elements*, Chapter 97, §3; emphasis in the original).

Tesnière goes on to lay out his understanding of verbal valency in the meanwhile classic division of verbs into avalent verbs (*Elements*, Chapter 98), monovalent or intransitive verbs (*Elements*, Chapter 99), bivalent or transitive verbs (*Elements*, Chapter 100) and trivalent verbs (*Elements*, Chapter 107). This division is based on Tesnière's classification of verbal dependents into actants and circumstants (*Elements*, Chapters 50 through 57). The latter are typically expressed as adverbials and are optional (*Elements*, Chapter 56, §2-3), it is the former that are relevant for this discussion: they are, in all but name, what UD v1 terms core nominal dependents.

Actants are always nominals (*Elements*, Chapter 48, §6) and a verb can have three at most (*Elements*, Chapter 50, §5). The actants therefore come in three types, defined largely semantically as follows:

I. Actant I: "From a semantic point of view, the first actant is the one that performs the action." In traditional terms, this is the subject (*Elements*, Chapter 51, §6-7).
II. Actant II: "From a semantic point of view, the second actant is the one that bears the action" In traditional terms, this is the direct object (*Elements*, Chapter 51, §9-10).
III. Actant III: "From the semantic point of view, the third actant is what benefits or takes detriment from the action" (*Elements*, Chapter 51, §19). Tesnière describes it as "once known in traditional grammar under the name indirect complement, a designation that has recently been replaced ... by complement of attribution" (*El-*

ements), Chapter 51, §19).[3] From the examples given, it is obvious that this is the traditional indirect object.

Much has been done in the years since the publication of *Elements* to refine our understanding of valency, its place in dependency grammar and its theoretical aspects (cf. the overview Ágel et al. 2003), as well as in terms of its practical application to the description of a particular language (e.g. Welk 2011 or Herbst et al. 2004). It is the latter that is of particular relevance here, as such works elaborate on Tesnière's limited classification of actants and provide guidelines on the analysis of verbal valency frames. And while those guidelines will of course not be applicable cross-linguistically tout court, they may nevertheless serve as a useful starting point. One recent project in particular proved to be of great use in the task of analyzing the valency of Maltese verbs and establishing a clear distinction between core and non-core verbal arguments. In what follows, I will briefly outline the project in question, its approach to classifying verbal arguments and my adaptation of it to Maltese.

6.4.4.2.2 VALLEX

VALLEX (Lopatková et al. 2017), meanwhile in version 3.0, is a valency frame dictionary of Czech verbs based on the Functional Generative Description (FGD, see chapter 2, section 2.4.2); the same framework serves as the theoretical foundation for the Prague Dependency Treebank (Bejček et al. 2013) and the description of the Czech syntax based on it (Panevová et al. 2014). In the course of the compilation of VALLEX, much thought was given to the issues of identification and classification of verbal dependents. As with similar projects (e.g. Herbst et al. 2004), the decisions made were based on empirical investigation of the phenomena in question (Lopatková et al. 2017: 18) and great care was taken to provide a clear and operational (i.e. easily testable, Lopatková et al. 2017: 21-22) justification. This, along with the project's well-organized nature and success (as demonstrated by Panevová et al.'s 2014 groundbreaking description of Czech syntax and the expansion of VALLEX to Arabic by Bielický 2015), makes VALLEX an ideal starting point for the kind of analysis of verbal valency attempted here, my general aversion towards linguistic frameworks notwithstanding.

Valency in VALLEX is seen as a property of the tectogrammatic level (see Chapter 2, section 2.4.1.2) and is based on two types of distinction: actants vs. non-actants and obligatory dependents vs. optional dependents (Lopatková et al. 2017: 17). A short list of conditions (adapted from Lopatková et al. 2017: 18 and the more general formulation in Panevová et al. 2014: 40) is used to make the distinction:

I. Discounting coordination and apposition, a specific verbal dependent can appear with a single verb instance (i.e. in a single clause) only once

3 See the translators' note in Tesnière 2015: 104.

II. A particular verbal dependent can modify
   a. any verb
   b. a limited (potentially enumerable) set of verbs

Verbal dependents that satisfy condition I and at the same time satisfy condition IIb are considered actants, i.e. core dependents. Table 7.6 contains a list of verbal dependents classified as actants in VALLEX (Panevová et al. 2014: 40-50) along with their semantic characteristics (Lopatková et al. 2017: 20-21).

Actants	Semantic characteristics
ACT (actor)	Agent, Bearer of action or state, Causer, Experiencer
PAT (patiens)	Affected object
EFF (effect)	Result of action
ADDR (addressee)	Addresee
ORIG (origin)	Originator of action or state

Tab. 6.6: Actants (core dependents) in VALLEX

As Lopatková et al. note, "the first two actants, ACT and PAT, are identified syntactically, while for the other actants, semantic criteria is used" ("V zásadě platí, že první dva aktanty, ACT a PAT, jsou určovány syntakticky, zatímco pro určení dalších aktantů se zohledňují sémantická kritéria"; Lopatková et al. 2017: 19). This is a compromise solution, designed to avoid the two extreme positions often encountered in such situations, i.e. decisions based purely on the semantics of the verb and its dependents on one hand and decisions based purely on their syntax on the other. The latter position is the one that both Borg and Azzopardi-Alexander and the UD v1 guidelines take and it is thus the problem that is being solved here.

In VALLEX, verbal dependents that do not satisfy condition I are considered free dependents (Tesnière's circumstants). These are, as Lopatková et al. (2017: 17) note, typically adverbials of which a single verb can govern multiple and the only limitation is their semantic compatibility. Table 6.7 (adapted from Panevová et al. 2014: 28-29, 54-77) lists the types of free dependents used in VALLEX, classified by their semantics.

MANNER	PLACE	CAUSALITY
ACMP (accompaniment)	DIR1 (from)	AIM (aim)
BEN (beneficiary)	DIR2 (which way)	CAUS (cause)
CONTRD (contradiction)	DIR3 (to)	CNCS (concession)
CPR (comparison)	LOC (where)	COND condition
CRIT (criterion)		
DIFF (difference)	TIME	OTHER
EXT (extent)	TFHL (for who long)	ATT (attitude)
HER (heritage)	TFRWH (from when)	COMPL (complement)
MANN (manner)	THL (how long)	INTF (expletive)
MEANS (means)	THO (how often)	INTT (intention)
REG (regard)	TOWH (to when)	MOD (modal)
RESL (result)	TPAR (parallel)	
SUBS (substitution)	TSIN (since when)	
	TTILL (till)	
	TWHEN (when)	

Tab. 6.7: Free dependents in VALLEX

In addition to the two types of verbal dependents which correspond to Tesnière's division into actants and circumstants, VALLEX established a third category, the so-called quasi-valency dependents (Panevová et al. 2014: 50-54). These are verbal dependents that share characteristics of both actants and free dependents and are largely specific for Czech.

6.4.4.2.3 Core arguments in MUDTv1 and criteria for their identification

In finding a solution to the problem of classifying verbal arguments in Maltese, two steps were involved: first, the definition of the straightforward core dependents – nsubj, nsubjpass, dobj and iobj. For that purpose, I applied the same principles as those used in VALLEX for the identification of actants ACT and PAT, i.e. syntactic criteria, namely:

I. nsubj/nsubjpass are defined primarily as:
   i. the verbal dependent that agrees with the verb in gender, number and person
II. dobj/iobj are defined primarily as:
   i. verbal dependents that bear the LIL/LIL_DEF case markers
   ii. verbal dependents not bearing the case markers cited above or prepositions that are replaceable or can be co-referential with the respective member of either of the verbal clitic sets

These definitions are of course overly broad and thus problematic, especially for the subjects; the finer points will be addressed in the respective entries below.

The second step then entailed doing away with the all-encompassing category of nmod which covers both verbal (clausal) dependents and nominal dependents. For the

purposes of MUDTv1, the functional load of nmod was divided between the following relations, listed in the decreasing order of their straightforwardness:

I. Nominal dependents
   i. nmod:poss
   ii. nmod
II. Verbal dependents
   i. nmod:agent
   ii. nmod:advmod
   iii. nmod:obj

The most straightfoward cases were dealt with first, starting with the nominal dependents: the relation nmod:poss (see section 6.4.4.9.2) was established to cover possessive constructions, both the analytic ones using *ta'* (Borg and Azzopardi-Alexander 1997: 76), as well as the construct state (Borg and Azzopardi-Alexander 1997: 71). This was inspired largely by the way Hebrew treats possessive constructions (which was then adopted by the UD v2 guidelines, cf. Nivre, Ginter et al. 2016), although what is traditionally termed status constructus is assigned a separate label in Hebrew. Every other noun-headed phrase (including prepositional phrases) that does not qualify as any other relation (e.g. conj or appos) would then be assigned the label nmod (see section 6.4.4.9.1).

Turning to verbal dependents, the relation nmod:agent (which still counts among the more straightforward ones) was established for the agent in passive clauses, invariably introduced by the preposition *min* "from" (see section 6.4.4.3.6). This is a solution with a precedent in the Swedish UD v1 treebank and the Romanian UD v2 treebank (Nivre, Agić et al. 2017). A variant, obl:agent, is used systematically in a number of UD v2 treebanks (Nivre, Ginter et al. 2016), including Italian where it is employed for the passive agent in a construction identical to that of the Maltese dynamic passive (Maiden and Robustelli 2007: 284-285).

As for the two remaining relations, the nmod:advmod relation was then split off of nmod to be used for adverbials, which are either prepositional phrases or noun phrases (see section 6.4.4.5.1 for examples). The definition of an adverbial was that used by VALLEX for free dependents (see Table 6.7 above), with some extensions as discussed in the entry for nmod:advmod. And this brings us back the problem of prepositional phrases like those introduced by *bi* discussed in section 6.4.4.2.1 above: how can we tell between non-core dependents (free dependents) realized as prepositional phrases (i.e. nmod:advmod) and core dependents (actants) also realized as prepositional phrases?

To do that, I employed the three-part criteria used for the purposes of VALLEX described above, albeit slightly modified. The modifications were motivated by issues of obligatoriness: in VALLEX, this is where the so-called dialogue test (Sgall et al. 1980: 46, Lopatková et al. 2017: 22) comes in. This test uses the assumption that there is a dif-

ference between asking about information already provided and asking for additional information to be provided. To give a trivial example:

Dialogue 1:
A: John came.
B: Who came?
A: *I don't know.

Dialogue 2:
A: John came.
B: When?
A: I don't know.

The "I don't know" answer provided in Dialogue 2 is nonsensical (outside of situations where A's original statement was overheard or misheard): the information B inquires about has already been given and the constituent that provided it is therefore obligatory. In Dialogue 2, however, the question B asks is fully justified, as it was not provided in the conversation.

In the VALLEX analysis of verbal valency frames, the primary function of the test is to confirm the obligatory nature of a verbal dependent on the tectogrammatic level, even though it may be absent in the surface realization of the sentence. The distinction between the two levels is irrelevant for the purposes of my analysis which is not based on FGP, even though it might be helpful in scenarios like the definition of nsubj above: in Maltese (as in Czech), the subject can be realized solely through verbal affixes and so in most contexts, an actual nominal nsubj is not obligatory. Such appeal to semantics (which is what this is) might provide a consistent way of defining subjects, but it does not help with the other major problem one is typically faced with when analyzing verbal valency, the "one verb or many verbs" issue and the related problem of diathesis (cf. Perini 2015: 12-15, 17-20).[4] By way of example, consider the two uses of the Maltese verb *ra* "to see" in (39) and (40):

(39) *Inħares bla ma nara.*
I look without COMP I see.
'I look without seeing.'

[BCv3: 1993 Immanuel Mifsud - Il-Ktieb tas-Sibt Filgħaxija]

4 Perini sees this as two different problems: the former is purely semantic, as with the English verb "get" and its various uses; the latter involves the optionality (or, in terms of Lopatková et al. 2017: 22, surface deletion) of verbal dependents, especially objects. To my mind, this is the same problem: "get" in "get home" and "get a raise" differ in both semantics and in their valency frame; same is true of "see" in "see well" and "see a way out".

(40) *Inħares: nara sema iswed.*
I look: I see sky black.
'I look: I see black sky.'

[BCv3: 1993 Immanuel Mifsud - Il-Ktieb tas-Sibt Filgħaxija]

The most conspicuous thing about this pair is the status of the object in the two contexts: one would expect that a verb of perception like *ra* would require one as it does in (40); after all, you always perceive (see, hear, smell, touch) something. So what happened to that object in (39)? Is it just deleted from the surface structure of the sentence (as Lopatková et al. 2017:22 and Perini 2015: 19 would have it)? Or are those two really two different verbs? Just consider their semantics: (40) denotes the process of visually perceiving something. In contrast, (39) merely denotes a person's capability or ability (to engage in said process) and as such, it cannot take an object. It would seem to be the case that these are indeed two different verbs, one monovalent (intransitive), one bivalent (transitive). But what is then their relationship to each other? Is the bivalent one the original and the other one the result of diathesis? And if, what implications does this have for the status of the object?

All those are questions that the dialogue test cannot help to answer and so I had to come up with my own quick and dirty definition of when a prepositional phrase is a core verbal dependent. It goes like this:

I. A core verbal dependent is a verbal dependent that
   a. is obligatory (as evidenced in *BCv3*); and
   b. is not an adverbial (with the list of free dependents in VALLEX, see Table 6.7 above, as a rough guideline); and
   c. can only appear once with a single verb instance (VALLEX condition I).

II. A verb may not take an obligatory dependent, but if it does (as evidenced in *BCv3*) and the dependent
   a. is of a specific type; and
   b. fulfills a particular semantic role (VALLEX actants, see Table 7.6); and
   c. can only appear once (VALLEX condition I)

   this dependent is considered a core dependent.

I thus ended up with a binary decision tree with two main branches. The first main branch is rather self-explanatory, if somewhat complicated in its second sub-branch which involves semantic analysis. Consider, for example, the imperfect verb *ddependa* "to depend" which in BCv3 primarily occurs in 3rd person imperfect (11,712 hits, 54.16 per million) and takes only one type of nominal dependent, an obligatory prepositional phrase introduced either by *min* "from" or *fuq* "on". Both are normally locative prepositions, yet in this context, the phrases they feature in can hardly be interpreted as either directional (VALLEX dependents DIR1, DIR2 and DIR3) or locative (VALLEX LOC). They could be interpreted as adverbials of origin, but the semantics of the verb do not support this, as there is no movement denoted in the verb and in any case, this would only

apply to those phrases introduced by *min*. Such a prepositional phrase will therefore be considered a core dependent and determining its actual semantic role will be left for further work.

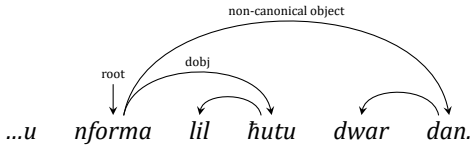
As for the second main branch, it is best illustrated by the example of the verb *nduna* "to notice" with a verbal dependent prepositional phrase introduced by *bi* "with". When I encountered such a construction, I queried *BCv3* for all forms of the verb to find that this verb typically takes no nominal dependents save for the subject, but if it does take other nominal dependents, they are overwhelmingly introduced by *bi* (subbranch IIa). Moreover, such dependents typically fulfill the semantic role of a direct object (patient, VALLEX PAT) in that they denote the person, object or phenomenon observed (subbranch IIb). As such, they cannot be considered either facultative or adverbials; for such dependents, I adopted the working designation "non-canonical objects".

The subbranch IIa is the best way I could come up to deal with VALLEX condition I (obligatoriness) and it works for subjects as well as object-like dependents. It also serves as a test for VALLEX condition IIb ("a particular verb dependent can modify a limited, potentially enumerable, set of verbs"), if in a roundabout way: corpus data is used to determine which dependents a particular verb takes and the decision tree is then applied to see if the prepositional phrase in question is a core dependent or a non-core dependent.

Based on the description above, one might have the suspicion that this type of analysis only investigates verbal government (Rektion) or a special subclass of verbs (so-called prepositional verbs, cf. Aquilina 1976: 67-80), but this is not the case: first, the primary purpose here is to determine which prepositional phrases can be considered core dependents and which are adverbials; this requires a more careful analysis than just "this verb takes dependents marked with this preposition". Secondly and more importantly, since Maltese verbs can take adverbials consisting of noun phrases unmarked for case (the UD v1 relation), such an investigation requires a careful analysis of all dependents of a particular verb, i.e. its entire valency frame, and cannot be reduced to the deciding whether a particular prepositional phrase is an adverbial or a non-core dependant. This path lead to a number of surprising and hitherto unexamined nooks in the study of Maltese, such as the classification of certain verbs as trivalent where the third core dependent (along with subject and direct object) is not a traditionally conceived indirect object marked by LIL/LIL_DEF. These include verbs like *tkellem* "to speak" or *nforma* "to inform" which both take a direct object and an obligatory (in terms discussed above) prepositional phrase introduced by *dwar* or *fuq* (41).

(41) *...u nforma lil ħutu dwar dan.*
...and he informed ACC brother.PL-his about this.M.

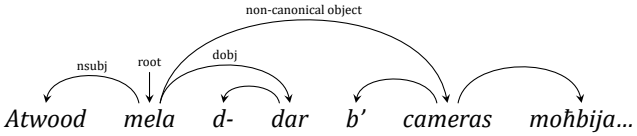
'... and he informed his brothers about this.'



[MUDTv1: 02_02J01]

These also include verbs like *mela* "to fill" which also take a direct object and an obligatory prepositional phrase introduced by *bi*; and while in the case of (41), the non-canonical object can conceivably be interpreted as fulfilling the same semantic role as direct objects (VALLEX actant PAT), this is definitely not the case for *mela* as in (42):

(42) *Atwood mela d- dar b' cameras moħbija...*
Atwood he filled DEF house with cameras hidden...
'Atwod filled the house with hidden cameras...'

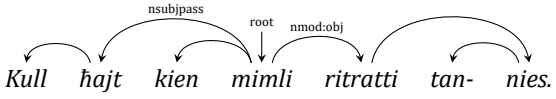


[BCv3: newspaper]

Here the semantics do not match the traditional view of direct and indirect objects: the non-canonical object *b'cameras* does not denote an entity something can be given, addressed or directed to; and even if one were to stretch the definition of those terms to argue it does, the house that is the direct object (patient) certainly cannot be given, addressed or directed to the cameras. In this clause, the non-canonical object denotes the result or effect of the action expressed by the verb, i.e. the VALLEX actant EFF.

This picture is further complicated by related phenomena, such as the passive diathesis (*Elements*, Chapter 102) of the verb *mela* illustrated in (43) below where its dobj turns into nsubjpass and its non-canonical object is a nominal unmarked for case; such a construction is not attested in *BCv3* for this verb in its active form.

(43) *Kull ħajt kien mimli ritratti tan- nies.*
every wall it was filled picture-PL GEN-DEF people.
'Every wall was filled with pictures of people.'



[MUDTv1: 46_02F08]

The detailed analysis of phenomena like diathesis is well beyond the scope of this work. Suffice it to say that the conceptual analysis elaborated on here, along with the principles and rules of UD v1, can be used to make quick and consistent annotation decisions in both (42) and (43), as well as other similar constructions.

In all of these, the newly established label nmod:obj is used for the non-canonical object in MUDTv1, as in (43). This label is the final one split off of the original UD v1 verbal dependent nmod and named to be consistent with the other two, nmod:agent and nmod:advmod. Unlike the latter two, the label dobj could have been used straight away, but I decided against it for one simple practical reason: if my analysis should turn out to be wrong (either wholesale, which I doubt, or in particular cases, which is nigh certain), this way, it is very easy to correct.

This concludes the discussion of the fundamental principles of analysis of verbal valency and the classification of nominal dependents of predicates in MUDTv1. Needless to say, this is far from the last word on the subject and while I am confident in the principles discussed above, some decisions may very well be questioned, especially when it comes to the distinction between adverbials and (non-canonical) objects which can be blurry under the best of conditions (cf. Jelínek 2015: 25). Only future detailed work on the valency of Maltese verbs can tell.

6.4.4.3 Core arguments: Nominals

6.4.4.3.1 Nominal subject: nsubj

This relation is used for the nominal subject of verbal (1), copular (9) and existential (31) clauses. A nsubj is primarily defined as the nominal dependent that agrees with the predicate in

i. gender, number and person (most verbal and all type (iv) copular clauses),
ii. gender and number (type (ii) and type (iii) copular clauses)

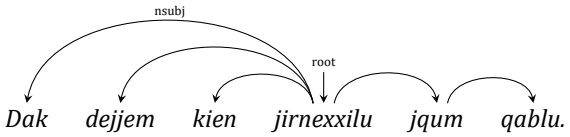
and as the only obligatory nominal dependent of the predicate for type (i) copular clauses and main existential clauses.[5]

For verbs (i.e. tokens tagged VERB), this agreement is invariably expressed through the use of subject affixes (the situation is somewhat more complicated with VERB_PSEU, for which see Peterson 2009). There are, however, verbs which mark this agreement using other suffixes, either in addition to or in place of subject affixes. These verbs fall largely into two categories: those that use additional suffixes and are typically reflexive, like *ħass* "to feel" which combines with direct object clitics. The other category contains verbs which are in essence impersonal, like *rnexxa* "to succeed", when they occur in their 3rd person singular masculine forms to which they take dative clitics (indi-

5 In dependent existential clauses, mostly acl and advcl, the subject or pivot is not mandatory, and in some cases its presence may even be ungrammatical. The former also goes for conj clauses.

rect object suffixes), typically to mark the undergoer or bearer or action (cf. Lopatková et al. 2017: 20). When they do occur with a nominal dependent, this is also annotated nsubj (44).

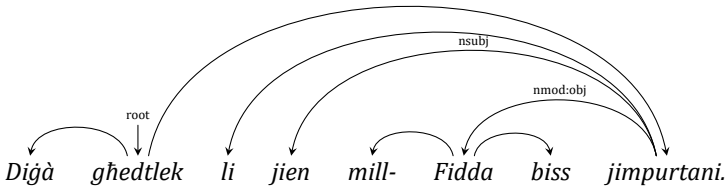
(44) *Dak dejjem kien jirnexxilu jqum qablu.*
that.M always he was he succeeded-DAT.3SG he rises before him
'That one always managed to get up before him.'



[BCv3: 2008 Loranne Vella Simon Bartolo-Wied Wirdien (Fiddien II)]

This also applies to a small class of verbs like *impurta* "to be of import, to care about" which mark ACT by a direct object clitic and the PAT either by a noun phrase unmarked for case or by a prepositional phrase introduced by *min-* "from". If the ACT is also realized by a noun phrase, it is annotated nsubj (45).

(45) *Diġà għedtlek li jien mill- Fidda biss jimpurtani.*
already I told-DAT.2SG COMP I from-DEF Silver only he cares-ACC.1SG
'I already told you that I only care about the Silver.'



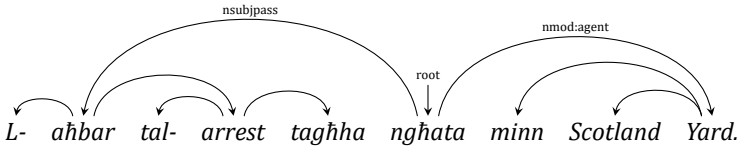
[BCv3: 2009 Loranne Vella Simon Bartolo - Il-Ġnien tad-Dmugħ (Fiddien III)]

In a small number of very specific cases, a nsubj can take a PREP or GEN/GEN_DEF as a dependent. For the former, this most commonly involves the preposition *madwar* "about (of number)" followed by a numeral; in such cases, *madwar* sensu stricto modifies the numeral. In the latter, these are occupational designations or membership designations, as in *tal-Lejber*, literally "GEN-DEF Labor", meaning "members of the Labor Party".

6.4.4.3.2 Nominal subject of a passive clause: nsubjpass

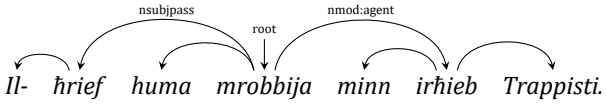
This relation is used for the nominal subject (i.e. the patient) in a passive clause (see section 6.4.4.1.2 above). It applies to passive clauses with both a passive VERB (46) and a PART_PASS (47) as the root.

(46) *L- aħbar tal- arrest tagħha ngħata minn Scotland Yard.*
DEF news GEN-DEF arrest her he was given from Scotland Yard.
'The news of her arrest was made public by the Scotland Yard.'



[BCv3: ilgensillum.2011-Awwissu-10.9366]

(47) *Il- ħrief huma mrobbija minn irħieb Trappisti.*
DEF lamb.PL they raised-F from monk.PL trappist-PL
'The lambs are raised by trappist monks.'



[BCv3: ilgensillum.2011-Frar-15.5933]

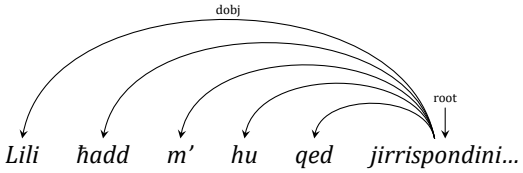
6.4.4.3.3 Direct object: dobj

This relation is used for the following verbal dependents:

I. Animate noun phrases introduced by LIL and LIL_DEF (cf. Borg and Azzopardi-Alexander 1997: 55) replaceable or co-referential with the direct object clitic set, including LIL_PRON when it acts as a patient satisfying the same condition (48).
II. Inanimate noun phrases unmarked for case in the semantic role of patient and replaceable or co-referential with the direct object clitic set (49).

(48) *Lili ħadd m' hu qed jirrispondini...*
ACC-1SG no one NEG NEG PROG he responds-ACC.1SG

'Me, no one responds to me.'

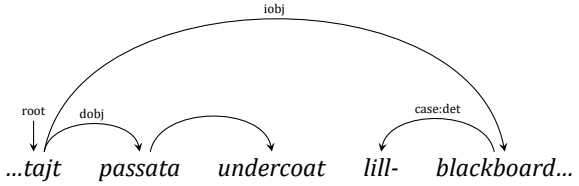


[MUDTv1: 38_02P06]

6.4.4.3.4 Indirect object: iobj

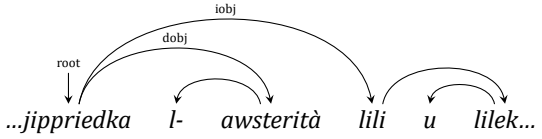
This relation is used for the indirect object, defined in both syntactic and semantic terms as a NOUN or noun phrase modified by LIL or LIL_DEF (49) or a one of the tokens tagged LIL_PRON (50) which denote addressee or benefactor of an action (VALLEX functor ADDR), as opposed to the patient or affectee where dobj is used.

(49) *...tajt passata undercoat lill- blackboard...*
I have coat undercoat DAT-DEF blackboard..
'...I gave the blackboard a layer of undercoat paint...'



[MUDTv1: 49_03F09]

(50) *...jippriedka l- awsterità lili u lilek...*
...he preaches DEF austerity DAT-1SG and DAT-2SG...
'... he preaches austerity to me and to you...'

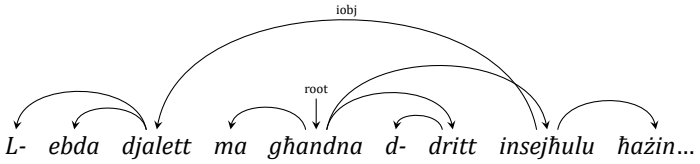


[MUDTv1: 14_01J02]

Additionally, the iobj relation is also used for core dependents of verbs that exhibit full conjugation, like *sejħa* "to call (by name), to refer to", that are co-referential with dative clitics, regardless of whether the verbal dependent in question is modified by

LIL/LIL_DEF or not (51) and for impersonal verbs like *ġara* in the sense of "to happen to" when the verbal dependent does take LIL/LIL_DEF. These dependents are more accurately described as patients and thus should perhaps be annotated as dobj, but here again the morphological criteria prevail.

(51) *L- ebda djalett ma għandna d- dritt insejħulu ħażin...*
DEF none dialect NEG we have DEF right we call-DAT.3SGM wrong ...
'We have no right to call any dialect wrong ...'



[MUDTv1: 52_03N10]

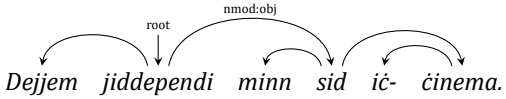
For impersonal verbs which take indirect object suffixes, but their dependents do not take LIL/LIL_DEF marker, see section 6.4.4.3.1 above.

6.4.4.3.5 Non-canonical object: nmod:obj

The nmod:obj is used for the following obligatory (see definition in 6.4.4.2.3) verbal arguments realized as nominals:

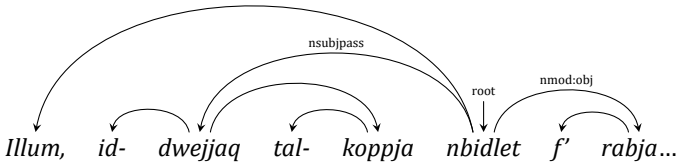
I. Obligatory arguments realized by means of prepositional phrases, regardless of how many and what other arguments the verb has (52).
II. Obligatory arguments realized by means of prepositional phrases that alternate with noun phrases (including clitics) normally annotated as dobj in the sort of variation described in section 6.4.4.2.1.
III. Obligatory arguments indicating the outcome or effect (the VALLEX actant EFF) or origin (VALLEX actant ORIG), regardless of whether they are realized as noun phrases (43) or prepositional phrases (53) and regardless of how many and what other arguments the verb has.

(52) *Dejjem jiddependi minn sid iċ- ċinema.*
always he depends from owner DEF cinema.
'It always depends on the owner of the cinema.'



[BCv3: illum.2007-10-07.t6]

(53) *Illum, id- dwejjaq tal- koppja nbidlet f' rabja...*
today, DEF sadness GEN-DEF couple it was changed in anger...
'Today, the couple's sadness turned into anger...'



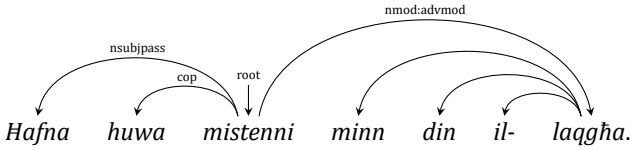
[MUDTv1: 05_05J01]

6.4.4.3.6 Nominal modifier - passive agent: nmod:agent

This relation is used for the optional agent noun phrase in passive clauses, invariably introduced by the preposition *minn*, regardless of the root (see examples (3)-(7) above).

Here once again the semantics and valency of the verb must be carefully considered when analyzing such clauses, as some passive constructions can also feature dependents consisting of a prepositional phrase introduced by *minn* which are not agents:

(54) *Hafna huwa mistenni minn din il- laqgħa.*
much he expected from this.F DEF meeting.
'Much is expected from this meeting.'



[MUDTv1: maltarightnow.2009-5-18.54-99812382]

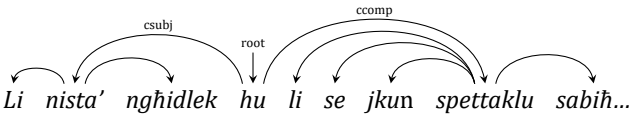
In this case, the prepositional phrase introduced by *minn* does not indicate the agent, as only persons and other beings and entities high on animacy scale can expect; instead, it indicates the source (direction from) and as such, it is annotated as nmod:advmod.

6.4.4.4 Core arguments: Clauses

6.4.4.4.1 Clausal subject: csubj

Clausal subjects are subjects of active clauses that are themselves clauses (cf. Borg and Azzopardi-Alexander 1999: 30). They are typically introduced by the complementizer/subordinator *li* and such clauses are typically copular and often feature a ccomp as the predicate (see 6.4.4.4.4 below), as in (55).

(55) *Li nista' ngħidlek hu li se jkun spettaklu sabiħ...*
COMP I can I tell you he COMP FUT he is show pretty...
'What I can tell you is that it will be a beautiful show...'

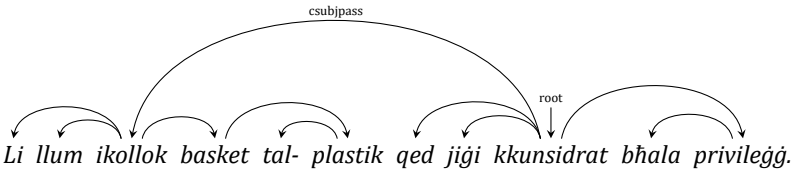


[MUDTv1: 23b_04J03]

6.4.4.4.2 Clausal subject of a passive clause: csubjpass

This relation is used for the equivalent of csubj in passive clauses of all types, as in (56).

(56) *Li llum ikollok basket tal- plastik qed jiġi kkunsidrat*
that today you have bag GEN-DEF plastic PROG he comes considered
bħala privileġġ.
as privilege.
'To have a plastic bag today is considered a privilege.'



[BCv3: 2012 Raymond Muscat - Naħqa Ta' Hmar]

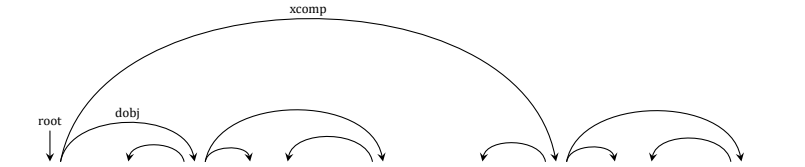
Such sentences, however, are not represented in MUDTv1.

6.4.4.4.3 Open clausal complement: xcomp

A complement clause is a clause that has the same function and status as an object (cf. Borg and Fabri 2016: 425); as such, it is considered a core dependent. In UD v1, an xcomp is a complement clause that inherits its subject from its governor or from its superordinate clause (Nivre, Ginter et al. 2014): in (57), the subject of the xcomp is the object of the main clause; in (58), the subject of the xcomp is the same as the subject of the main clause. In both cases, however, the xcomp attaches directly to the predicate.

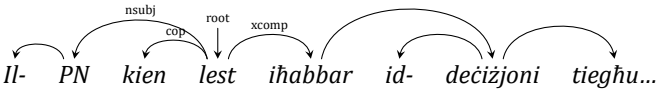
(57) *Nappella lill- awturi u lill- kompożituri biex*
I encourage ACC-DEF author-PL and ACC-DEF composer-PL in order to
jimirħu u ma jistaġnawx.
they expand and NEG they stagnate-NEG

'I encourage authors and composers to expand and not to stagnate.'



[MUDTv1: 23b_04J03]

(58) *Il- PN kien lest iħabbar id- deċiżjoni tiegħu...*
DEF PN he was ready he announces DEF decision his...
'The Nationalist Party was ready to announce their decision.'

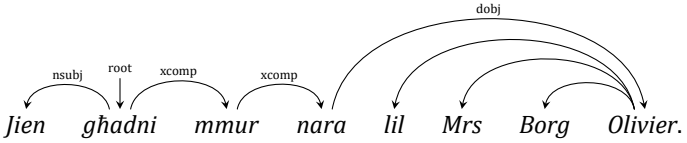


[MUDTv1: 04_04J01]

In Maltese, such clauses are typically introduced by the complementizer *li* (this is the "noun clause" described in Borg and Azzopardi-Alexander 1997: 30-33) and the subordinators *biex* and *jekk* (Borg and Fabri 2016: 421), but they can also bear no marker at all; in MUDTv1, the presence or absence of the marker therefore plays no role, only the syntactic relationship as described above does.

The second use of this relation in MUDTv1 is to annotate verbs connected in a verbal chain (see section 6.4.4.8.1 on aux), as in (59).

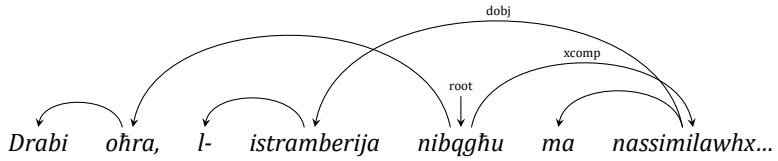
(59) *Jien għadni mmur nara lil Mrs Borg Olivier.*
I I still do I go I see ACC Mrs Borg Olivier.
'I still go see Mrs. Borg Olivier.'



[MUDTv1: 22_02J03]

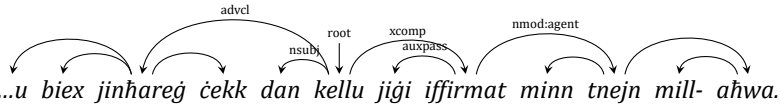
In this context, the application of UD v1 annotation Rule V (see section 6.4.3) becomes a little problematic. Consider the direct object *l-istramberija* in (60) and the subject in (61):

(60) *Drabi oħra, l- istramberija nibqgħu ma*
time-PL other-F, DEF strangeness we remain NEG
nassimilawhx…
we assimilate-ACC.3SF-NEG
'Other times, we are unable to assimilate the strangeness...'



[MUDTv1: 50_01N10]

(61) *…u biex jinħareġ ċekk dan kellu jiġi iffirmat*
...and in order to he is issued check this.M he had to he comes signed
minn tnejn mill- aħwa.
from two from-DEF brother.PL
'... and in order for a check to be issued, it had to be signed by two of the brothers.'



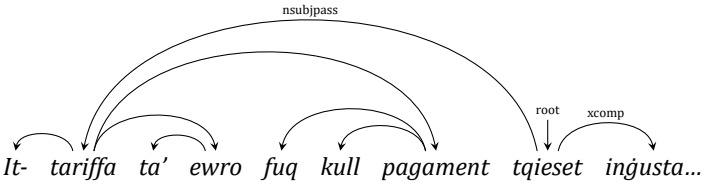
[MUDTv1: 02_02J01]

As noted in section 6.4.4.8.1 (see also Stolz 2009: 138 and Fabri and Borg 2017: 70), the verbs in a verbal chain all share the same subject; ipso facto, it is also the subject of the last verb in the chain which is, in accordance to the Many Auxiliary Theory, the lexical verb. The problem is that in (61), the lexical verb (passive participle, in this case) is passive, which would normally necessitate applying the nsubjpass label to *dan*. On the other hand, *dan* is also the subject of the first verb in the chain which is an intrasitive one and thus cannot be made passive. This is an unfortunate downside to adopting the One(-ish) Auxiliary Theory: if *baqa'* were treated as an auxiliary, this problem would not exist, but since it is treated as a separate clause, it needs to be dealt with. Attaching the dobj *l-istramberija* in (60) to *nassimilawhx* is an obvious choice, since this is the only bivalent verb in this sentence; the resulting non-projective dependency is an acceptable tradeoff. It now needs to be decided what to do with the subject: to attach it to *iffirmat* would not only create a non-projective dependency, but it also wouldn't be accurate as *dan* is, after all, the subject of *kellu*, but it would also leave *kellu* hanging (no pun intended). The solution in (61) is therefore adopted throughout: in verbal chains

clauses featuring a nsubj and an dobj or nmod:obj, the former attach to the first or last verb in the chain, the latter attach to the verb into whose valency frame they fall.

The third typical use of the xcomp relation is for the so-called secondary predicatives, such as the one in example (62) below.

(62) *It- tariffa ta' ewro fuq kull pagament tqieset inġusta...*
DEF charge GEN Euro on every payment she was perceived unjust-F...
'The charge of one Euro on every payment was perceived as injust.'



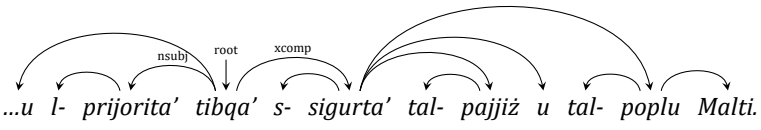
[MUDTv1: 07_07J01]

These are structures where there are two predicates with the same subject rolled into one and the clause can be paraphrased (if somewhat clunkily) by a combination of a main clause and a subordinate clause with the same subject referent; in this case, something along the lines of *It- tariffa ta' ewro fuq kull pagament tqieset* ***li kienet*** *inġusta* "The charge of one Euro was perceived **that it was** unjust." In such clauses, the second predicate attaches to the first one as xcomp. These sentences (partially equivalent to Borg and Azzopardi-Alexander's "adjectivalized noun clauses", 1997: 34) typically involve verbs of perception such as *tqies* "to be felt" (62), *deher* "to appear" and *ra* in the sense of "to view, to consider" (cf. Borg and Azzopardi-Alexander's "adjectivalized noun clause", 1997: 34-35).

By extension, this label is also used in case of verbs that denote apparent change (or lack thereof) in identity, property or state; namely it is applied to their obligatory nominal or adjectival dependents that denote said identity, property or state. These include, but are not limited to: *baqa'* "to stay, to remain" (63), *ġie* in the sense of "to become", *ħareġ* in the sense of "to become, to end up", *insab/instab* "to be found", *laħaq* "to attain a position of X", *qagħad* "to stay", *sar* "to become", *spiċċa* "to end up as something" and *żamm* "to keep, to remain".

(63) *...u l- prijorita' tibqa' s- sigurta' tal- pajjiż u*
...and DEF priority she remains DEF security GEN-DEF country and
tal- poplu Malti.
GEN-DEF people Maltese.

'... and the priority remains the security of the country and of the Maltese people.'

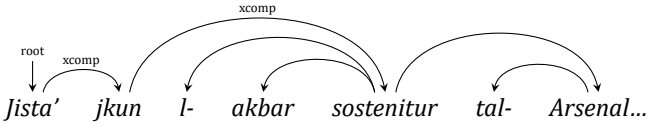


[MUDTv1: 13_13J01]

This decision resolves the conundrum hinted at by Borg and Azzopardi-Alexander (1997: 53) regarding copular predication: typically, copulas are used to express identity and attribution, but the verbs listed above could conceivably be viewed as doing the same which would complicate the classification of copular clauses. Borg and Azzopardi-Alexander only cite *sar* and *insab*, noting the very strong case in favor of considering the latter a full copula, yet ultimately deciding against it (but see the reversal of that position in Borg and Spagnol 2015). Since the same decision had been made in MUDTv1, it had to be determined what to do with the dependents of these verbs which are invariably obligatory and which no longer could be classified as copular predicates. Based on its use for secondary predicates which are very similar in their syntactic behavior, xcomp is the perfect choice for such constructions.

By further extension, the xcomp relation is used for the predicates of copular clauses that are the last link in a verbal chain, as in (64).

(64) *Jista' jkun l- akbar sostenitur tal- Arsenal...*
he can he is DEF biggest supporter GEN-DEF Arsenal...
'He might be the biggest fan of Arsenal...'



[MUDTv1: 18_05J02]

This is a clunky and ugly solution, but it is consistent (and thus easy to analyze and fix) and it is in line with the preferred UD v1 and UD v2 solution in a similar context (see section 6.4.4.4.4 below).

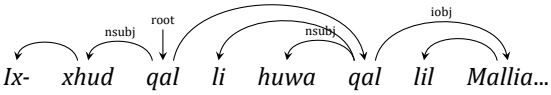
6.4.4.4.4 Clausal complement: ccomp

In contrast to xcomp, ccomp is a complement to a verb, an adjective or an adverb, which has its own subject. Determining this requires first and foremost a a thorough analysis

of the syntax and the semantics of both the candidate ccomp clause and the higher clause(s), as well as their context.

In MUDTv1, one additional criterion was applied: if the subject of in the candidate verbal clause is identical to one in the higher clause, but it is overt, the candidate clause will be annotated as ccomp. The logic behind this is that an overt subject is not required in Maltese verbal clauses and so when one is supplied as in a stereotypical example of a main clause with a ccomp in (65), it serves to ensure the correct interpretation of the subject in the dependent clause; this is the purpose of ccomp in UD (cf. the respective entry in Nivre, Ginter et al. 2014).

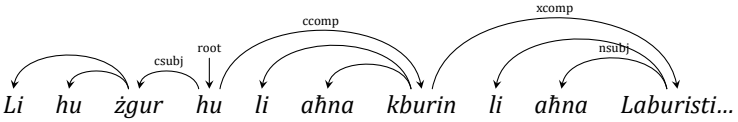
(65) *Ix- xhud qal li huwa qal lil Mallia...*
DEF witness he said COMP he he said to Mallia...
'The witness said that he said to Mallia...'



[MUDTv1: 02_02J01]

This does not apply to copular complement clauses which, with the exception of type (iv) copular clauses featuring KIEN, require a subject to be copular clauses. To illustrate, consider (66):

(66) *Li hu żgur hu li aħna kburin li aħna Laburisti...*
COMP he certain he COMP we proud-PL COMP we Laborist-PL...
'What is certain is that we are proud that we are Laborists...'

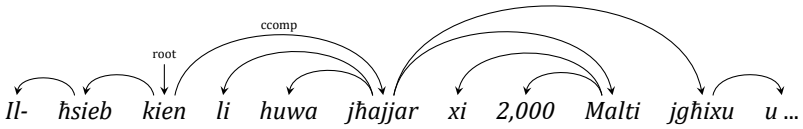


[MUDTv1: 22_02J03]

The place of the nsubj *aħna* could be taken by the respective form of KIEN, *nkunu*, which would be a clear-cut case of the complement clause inheriting a subject from its superordinate clause. On this logic, xcomp is used for those copular complement clauses that feature a nominal subject unless, of course, said subject is different from that in the higher clause.

And finally, ccomp is also used in copular sentences where the predicate is a clause, such as example (67) below:

(67) *Il- ħsieb kien li huwa jħajjar xi 2,000 Malti jgħixu u*
DEF thought was COMP he entices some 2,000 Maltese they live and
jaħdmu ġo Margo.
they work inside Margo.
'The idea was that he would entice some 2,000 Maltese to come to live and work in Margo.'



[MUDTv1: 52_03N10]

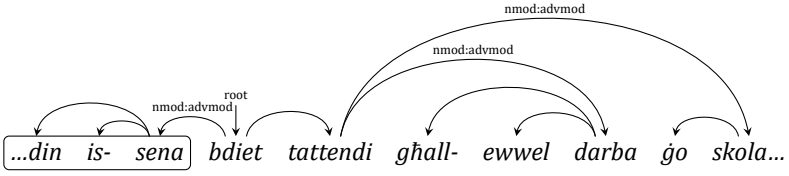
In this context, the copula (of any type) is treated as a head and the root of the clause, contrary to its normal status as a dependent of the predicate. As Nivre, Ginter et al. (2014) note, this is "a somewhat inconsistent and ugly feature of the current UD" i.e. UD v1. An update to this feature is planned, but as of UD v2, this is still the preferred way of annotating these structures, except the justification is "to preserve the integrity of clause boundaries and prevent one predicate to be assigned two subjects" (Nivre, Ginter et al. 2016).

6.4.4.5 Non-core dependents: Nominals

6.4.4.5.1 Nominal modifier - adverbial: nmod:advmod

This relation is primarily used for prepositional phrases (including PREP_PRON) which act as adverbials, i.e. facultative dependents of the predicate (of any type) that are not a nmod:obj or nmod:agent and that fulfill one of the semantic roles listed in Table 6.7 above (68).

(68) *...din is- sena bdiet tattendi għall- ewwel darba ġo skola...*
...this.F DEF year she began she attends on-DEF first time in school...
'... this year, she began to attend school for the first time.'



[MUDTv1: 39_01F08]

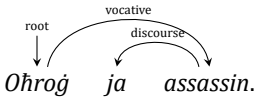
The list in Table 6.7 is the starting point for the analysis, not an exhaustive list; consequently, there are nmod:advmod that go beyond that set, such as prepositional phrases introduced by *bħal* "like, as" which denote a role or status.

In addition to prepositional phrases, nmod:advmod is also used for adverbial noun phrases that are unmarked for case (i.e. do not have a preposition as a dependent). These are either locatives or temporal adverbials as the highlighted catena in (68).

6.4.4.5.2 Vocative: vocative

This relation is used for nouns, personal pronouns and names denoting persons and person-like entities who are being addressed (69) or invoked. This includes tokens tagged as X_ABV or X_ENG, such as *Mr. Chairman* and *Mr. Speaker*, both ubiquitous in parliamentary texts.

(69) *Oħroġ ja assassin.*
leave voc murderer
'Leave, murderer.'



[MUDTv1: 47_01F09]

These words can take a limited set of dependents, like the vocative particle *ja* and its reduced form *j'* in the discourse relation to it (69) and tokens tagged X_ENG and X_ABV in foregin and name relations to it.

6.4.4.5.3 Expletive: expl

This relation is used for expletive subjects in non-copular verbless clauses; for examples, see (20) through (27) in section 6.4.4.1.4 above. As evident from the discussion there, expl is also used for KIEN in this role, as neither of the other relations applicable to KIEN (cop and aux) is really applicable here. On the other hand, PRON_PERS_NEG in examples like (28) will always be annotated as neg on morphological grounds.

6.4.4.5.4 Dislocated elements: dislocated

In the context of this work, this relation is somewhat problematic: first, the very concept of "dislocation" is laden with so much theoretical baggage and used in so many different meanings so as to be nearly useless. The UD v1 guidelines (Nivre, Ginter et al. 2014) are a good example of that: on one hand, they describe the purpose of this relation as annotating "fronted or postposed elements that do not fulfill the usual core grammatical relations of a sentence", adding that "[t]hese elements often appear to be in the periphery of the sentence, and may be separated off with a comma intonation".

At the same time, however, they note that dislocated is "used for fronted elements that introduce the topic of a sentence", but it should "not be used for a topic-marked noun that is also the subject of the sentence".

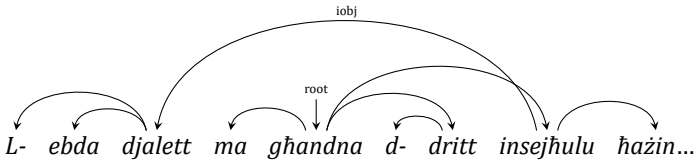
All of this leaves us with three broad criteria to define a dislocated element: a syntactic one ("do not fulfill the usual core grammatical relations"), a phonological one ("comma intonation") and a pragmatic one ("elements that introduce the topic"); the application is left to the compilers of individual treebanks. And here is where the problems begin: due to the written nature of the texts, the phonological criterion is out the window outright – commas are not a reliable indicator of changes in intonation or phonological breaks at the best of times, let alone in texts produced by such careless creatures as Maltese journalists. But even if they were, such suprasegmental phonological phenomena are not a sufficient condition for the word or phrase in question to qualify, hence the "often" in the definition above. For that, we must turn to the analysis of the syntactic goings on at either periphery. I attempted just such an analysis in my paper on object reduplication using multiple criteria (syntactic conectedness, iterativity and embedding, Čéplö 2014: 209-211). I arrived at the tentative conclusion that at least for some types of constructions (Hanging Topic and Clitic Left Dislocation), there seems to be a considerable degree of variation and ambiguity (Čéplö 2014: 212).

But even if a satisfactory definition and/or test could be found, there will still remain one large problem: it would be sharply at odds with the purpose of this work. I have set out to study Maltese constituent order without any theoretical preconceptions and there is no better example of such preconceptions than "dislocation". As I noted in the formulation of the research questions, the process I am following here requires that I first describe the variation and only then attempt to interpret and explain it. Describing something as "dislocation" involves the interpretation of data. And while the same is true of just about every annotation decision described here, the term "dislocation" involves a much larger conceptual apparatus than the one I am using here which consists of morphological analysis, semantic analysis, the general Principles and Rules of UD v1 and verbal valency.

For all these reasons, I originally decided not use this relation at all and instead apply Rule V of UD v1 annotation: in other words, whatever the position of a word or a phrase, its syntactic relation to its governor is annotated based on its place in the latter's valency frame, even if this results in non-projective dependencies. The iobj *l-ebda djalett* in (70) is a straightforward example of this.

(70) *L- ebda djalett ma għandna d- dritt insejħulu ħażin...*
DEF none dialect NEG we have DEF right we call it wrong ...

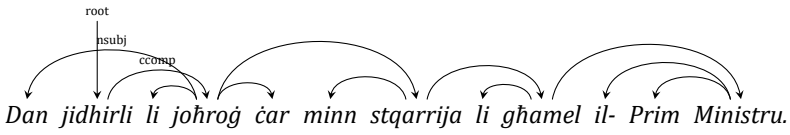
'We have no right to call any dialect wrong ...'



[MUDTv1: 52_03N10]

Soon, however, I came across cases that gave me pause, like the nsubj demonstrative pronoun *dan* in (71), which is separated from its verb by an entire (impersonal) clause.

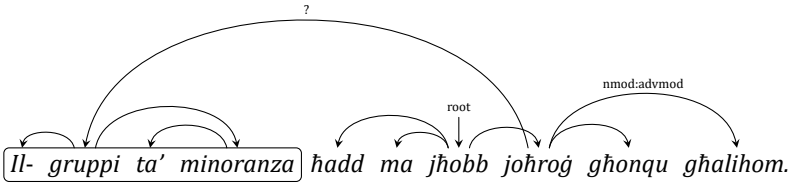
(71) *Dan jidhirli li joħroġ ċar minn stqarrija*
this.M he appears-DAT.1SG he comes out COMP clear from statement
li għamel il- Prim Ministru.
COMP he made DEF prime minister.
'This, it appears, comes out clear in the statement the Prime Minister made.'



[BCv3: ilgensillum.2012-Lulju-22.15685]

And then finally, sentences appeared in which a straightforward annotation of core verbal dependents was impossible to accomplish (72):

(72) *Il- gruppi ta' minoranza ħadd ma jħobb joħroġ għonqu*
DEF group-PL GEN minority no one NEG he likes he takes out his throat
għalihom.
on them.
'No one likes to say bad things about minority groups.'

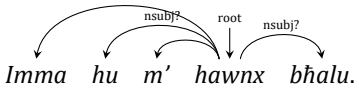


[MUDTv1: 22_02J03]

There are two problems with this sentence: first, the prepositional phrase *għalihom* "on them" and the noun phrase *Il-gruppi ta' minoranza* "minority groups" have the same

referent; secondly, the latter's function as an adverbial only becomes clear once the prepositional phrase is encountered. Whether the highlighted catena is dislocated or not is of little relevance at this moment; what is really important for the purposes of annotation is the fact that we have two constituents with the same referent. That is the real problem, especially if we consider that any constituent, like the subject *din* "this" in (8) above or the apparent subject *hu* "he" in (73) below, can be reduplicated in that way:

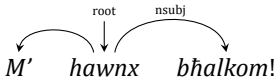
(73) *Imma hu m' hawnx bħalu.*
but he NEG EXIST-NEG like him.
'But he, there's no one like him.'



[BCv3: 2007 Mario Azzopardi - Alicia titkellem mill-Imwiet]

Example (73) is particularly interesting: it is an existential clause and in such clauses in Maltese (as represented in *BCv3*), the existential predicate HEMM only takes three arguments – the obligatory nsubj (or "pivot", cf. McNally 2011: 1833), an optional auxiliary and an optional coda (cf. Bentley 2015:2); in Maltese, the latter is typically an advmod, a nmod:advmod or a complement clause). The PREP_PRON *bħalu* above is neither and, more importantly it is not optional, as leaving it out result in an ungrammatical sentence or at the very least a change in meaning. Consequently, it can only be a subject (pivot); and in fact, this is shown by the existence of sentences such as (74):

(74) *M' hawnx bħalkom!*
NEG EXIST-NEG like you.PL.
'There's no one like you!'



[BCv3: darba-wahda-kien-hemm-teatru_july222013]

And so while I have no use for the dislocated relation in its original definition, I am using it in MUDTv1 to annotate duplicated dependents like the highlighted catena in (72) or the personal pronoun *hu* in (73); the decision on which one of the two will be labeled with its actual role and which one will be considered dislocated will be based on analyses such as the one comparing (73) (74). Whether or not those dependents labeled dislocated fulfill the definition of a dislocated dependent as per UD v1 is beside the point; here once again I am aiming not at full descriptive adequacy (which in this

case is not possible without the annotation of secondary relations), but at simplicity and consistency.

Now obviously, what I said about the adverbial in (72) and the subject in (73) is true of the noun phrase *l-ebda djalett* and the clitic *-lu* in *insejħulu* in (71); if it's not immediately obvious, it's only because my choice (forced as it may have been) not to split off clitics obscurs it. One would think that if I did split off the clitics and annotate their syntactic relationship to the verb, I would be faced with the same problem for all dobj and iobj which could only be resolved by annotating the lexical object as dislocated. That is not true: as I showed (Čéplö 2014: 209-211) and as Fabri notes (Fabri 1993: 145-146), there are many OV contexts in which resumptive clitics are not required. And so while some objects may be separated from the rest of the clause by a phonological break or may lack morphological or syntactic markers normally required, they still fulfill their role as a core dependent.

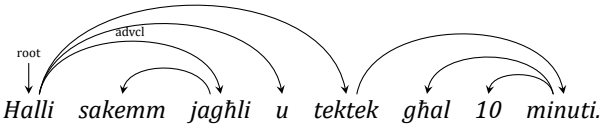
In light of all of this, I have adopted the following definition of a dislocated element: it is a dependent of the clause root that has the same referent and function as another dependent. This of course exludes verbal clitics which are not distinct words in MUDTv1. The question of what to do with them once split off will be left for MUTDv2; the preliminary solution I am considering is to establish a separate relation for the clitics, say, dobj:cl and iobj:cl or even annotate co-reference. This would be advantageous not only for the purposes of further analysis of constituent order, but also for the study of such phenomena as clitic doubling proper (Čéplö 2014: 218-220, Souag 2017), but will be left for future work.

6.4.4.6 Non-core dependents: Clauses

6.4.4.6.1 Adverbial clause: advcl

An adverbial clause is the clausal equivalents of advmod. In MUDTv1, they are indentified according to the classification provided by Borg and Azzopardi-Alexander (1997: 37-46), like the adverbial clause of time in (75).

(75) *Halli sakemm jagħli u tektek għal 10 minuti.*
leave until he boils and simmer on 10 minutes.
'Leave it until it boils and simmer for 10 minutes.'



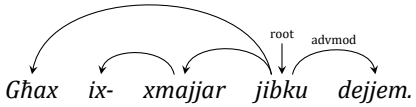
[MUDTv1: 54_01N11]

6.4.4.7 Non-core dependents: Modifier words

6.4.4.7.1 Adverbial modifier: advmod

This relation is used for tokens tagged ADV that modify predicates of any type, as (76) modifying a verbal predicate.

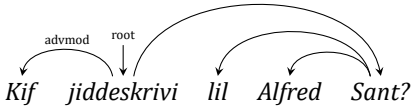
(76) *Għax ix- xmajjar jibku dejjem.*
because DEF river.PL they cry always.
'Because the rivers always cry.'



[MUDTv1: 46_02F08]

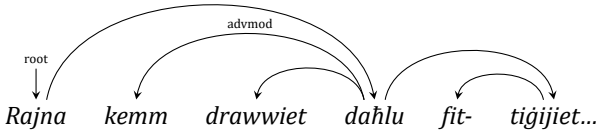
Additionally, this relation is used for those PRON_INT employed in asking about time, location, manner and count, whether in interrogative clauses (77) or in subordinate clauses (78).

(77) *Kif jiddeskrivi lil Alfred Sant?*
how he describes ACC Alfred Sant?
'How does he describe Alfred Sant?'



[MUDTv1: 22_02J03]

(78) *Rajna kemm drawwiet daħlu fit- tiġijiet...*
we saw how many custom-PL they entered in-DEF marriage-PL...
'We saw how many customs involved marriages...'

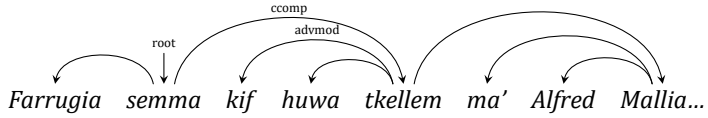


[MUDTv1: 53_04N10]

In subordinate clauses, PRON_INT tokens could also be considered subordinators and thus be annotated with the mark relation (see section 6.4.4.8.4 below); this is especially true of PRON_INT used in ccomp, as *kif* in (79) where it cannot conceivably be interpreted as an adverbial and is fully equivalent to (and replaceable with) *li*. Having

consulted selected UD v2 treebanks, I decided against that, if only for reasons of consistency and compatibility with the rest of UD treebanks.

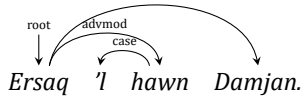
(79) *Farrugia semma kif huwa tkellem ma' Alfred Mallia...*
Farrugia mentioned how he he spoke with Alfred Mallia...
'Farrugia mentioned how he spoke to Alfred Mallia...'



[MUDTv1: 02_02J01]

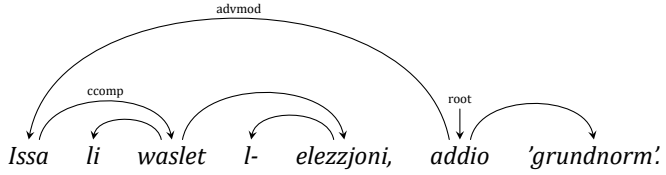
By their nature as modifiers, advmod dependents can only take a limited set of dependents of their own. These are primarily other adverbs and case markers in the case relation to it (80), but some cases, as with the adverb *issa* "now", this can be an entire ccomp as in (81).

(80) *Ersaq 'l hawn Damjan.*
approach to here Damjan.
'Come here, Damjan.'



[MUDTv1: 47_01F09]

(81) *Issa li waslet l- elezzjoni, addio 'grundnorm'.*
now COMP she arrived DEF election, goodbye 'grundnorm'.
'Now that the election is here, goodbye to *grundnorm*.'

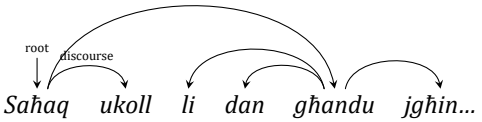


[MUDTv1: 14_01J02]

6.4.4.7.2 Discourse element: discourse

This class of dependents primarily contains various discourse markers tagged INT. Additionally, it is used for all focus particles, i.e. tokens tagged FOC as *ukoll* "also, as well" in (82).

(82) *Saħaq ukoll li dan għandu jgħin...*
he claimed also COMP this.M he has to he helps...
'He also claimed that this must also help...'



[MUDTv1: 10_10J01]

Here one should note the variability of focus particles scope: focus particles can modify nearly any word class and so be a dependent of not just the predicate, but also of any of its dependents. As such, they perhaps should not be included in a relation which is classified as a non-core clause root dependent, but then again, they would not be at home among nominal modifiers, either. So here is where they stay.

The verb *jiġifieri* "it means", normally tagged as such, is also assigned the discourse relation when used in its discourse function introducing a paratactic comment or appos.

6.4.4.8 Non-core dependents: Function words

6.4.4.8.1 Auxiliary verb: aux

The UD v1 standard defines an auxiliary verb as "a verb that accompanies the lexical verb of a verb phrase and expresses grammatical distinctions not carried by the lexical verb, such as person, number, tense, mood, aspect, and voice" (Nivre, Ginter et al. 2014), noting that in some languages, modal verbs fall into that category as well (see also Givón 2001a: 71). The problem with this definition is that it only works well within the tense-aspect-modality system of Standard Average European. The stereotypical auxiliaries of this definition are therefore those used with past participles in the typical SAE perfect (whether the *habeo*-perfect as in Germanic and Romance or the Slavic and Baltic *sum*-perfect) or those used with infinitives to form various types of future and modal structures. This doesn't translate very well to languages which do not possess infinitives or form tenses using participles; in fact, it doesn't even translate well between languages with different types of non-finite verbal forms. To give an example, the English UD v1 treebank classifies the following verbs as auxiliaries: "have", "do", "be", "will", "would" and "should" and the modals "can", "may" and "must". It does not, however, consider "want" an auxiliary, for obvious reasons: the aux verbs on the list depend on the bare

infinitive; "want" requires the presence of the "to" particle in its complements. As such, it is considered a full verb and a head of its own clause. And so whereas the UD v1 definition cited above is motivated morphologically (or rather by the absence of relevant verbal morphology, cf. "expresses grammatical distinctions not carried by the lexical verb") and semantically (the reference to "lexical verb"), the practical application of those guidelines in the English UD v1 treebank establishes a syntactic dimension to the analysis of the relationship between auxiliaries and lexical verbs.

Maltese does not possess SAE-style infinitives and it does not employ participles in the articulation of tense, mood and aspect, save for the present active participle of *qagħad* and the present active participle of *sar* (which are assigned their own part-of-speech tag and UD label due to their morphology and straightforward behavior when modifying verbs, see section 6.4.4.8.6 in this chapter). In other words, all verbs and pseudoverbs (with the exception of *għad*, *tantx* and HEMM) show at least the person and number distinction and the morphological criterion is thus out the window. Consequently, in deciding what should count as an auxiliary verb in Maltese, there are two routes to be taken.

The first route is laid out by Vanhove in her comprehensive analysis of the Maltese verbal system and the role of auxiliaries in it (Vanhove 1993: 101-329). Vanhove identifies a large set of "auxiliaries, verbal particles or preverbs" ("auxiliaires, particules verbales ou préverbes", Vanhove 1993: 101) where auxiliaries are defined in both functional and syntactic terms: as for the former, the auxiliaries serve to "instantiate the potentialities contained in the verb" ("représentent concrètement des potentialités contenues dans le verbe", Vanhove 1993: 101). In other words, they serve to express grammatical and semantic relationships the verb itself is incapable of expressing using other means, such as tense, aspect and mood, voice, concomitance, quantity (frequency, duration, intensity) and temporal relations (beginning, end) (Vanhove 1993: 101-102). In syntactic terms, Vanhove establishes a number of tests, such as the test of elimination (e.g. in *jaqbad jidħak* "he begins to laugh" can be reduced to *jidħak* "he laughs", but not to *jaqbad* because on its own it means "he grabs", thus establishing *jaqbad* as the auxiliary; Vanhove 1993: 102), the test of asyndecity (the entire phrase cannot be joined in coordination or subordination, thus establishing that the words are in a dependent relationship; Vanhove 1993: 103), the test of a single object (the entire phrase can only have one single object; Vanhove 1993: 103) and the test of a single subject (both verbs have the same subject; Vanhove 1993: 103). In addition to these definitions, however, Vanhove also employs a semantic criterion: in reference to the asyndecity test, she notes that in terms of syntactic structure, there is nothing to distinguish the auxiliary + verb construction *jaqbad jidħak* "he begins to laugh" and the verb + verbal complement construction *jidħol jidħak* "he enters laughing". The only way to distinguish the former from the latter is their semantics where in the former structure, the relationship of the two verbs is that of a unity while in the latter, it is a sum (Vanhove 1993: 102).

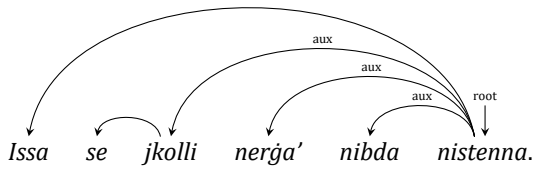
Using these criteria, Vanhove compiles a list of verbs (and pseudoverbs) that serve as auxiliaries in Maltese (Vanhove 1993: 153-330). These include (in alphabetical order): *baqa'* "to remain", *beda* "to begin", *fetaħ* "to open", *għad-* "to still be", *għand-/kell-/ikoll-* "to have", *għodd-* "to almost X", *ġie* "to come", *ħa* "to take", *ħabat* "to happen, to occur", *ħasel* "to occur", *sar* "to become", *issokta* "to continue", *jaf* "to know", *kien* "to be", *kompla* "to continue", *mess* "to touch", *qabad* "to grab; to set out to", *qabeż* "to jump", *qagħad* "to stay, to reside, to be in a place", *ra* "to see", *reġa'* "to return; to do again", *rema* "to throw", *ried* "to want", *safa* "to be in a state", *seta'* "to be able to", *telaq* "to leave", *wasal* "to arrive" and *żied* "to add".

Such a definition of an auxiliary makes perfect sense in the light of the phenomenon that Stolz terms "verbal chaining" (Stolz 2009, see also Fabri and Borg 2017).[6] Broadly defined, this is a phenomenon where "minimally two verbs [form a sequence] in one and the same utterance that is not interrupted by the insertion of subordinating conjunctions and whose members share the same subject" (Stolz 2009: 138), typically in a rigid linear sequence ordered by type with the lexical verb at the end (Stolz 2009: 150):

> Aux > (Aux) > Pseudo > (Pseudo) > TMA > Phasal > (Phasal) > Modal > Pseudo > (Pseudo) > Phasal > (Phasal) > TMA > Lex > (Lex)

Example (83) below contains a typical verbal chain:

(83) *Issa se jkolli nerġa' nibda nistenna.*
now FUT I will have to I return I begin I wait
'Now I will have to start again begin waiting.'



[BCv3: 1993 Immanuel Mifsud - Il-Ktieb tas-Sibt Filgħaxija]

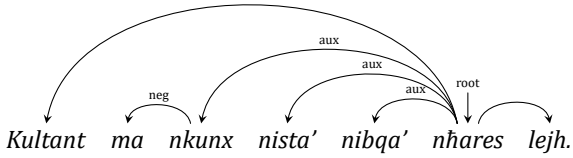
As Stolz notes, structures like the verbal chain in (83) are not examples of serial verb constructions, but rather a "combination of a number of elements each of which contributes in principled ways its share to the grammatical interpretation of the entire complex" (Stolz 2009: 175). This analysis is of course entirely consistent with Vanhove's view (which I will refer to as the Many Auxiliaries Theory) and is in fact what Maas (2009: 114) argues for as well. In semantic terms, examples like (83) consist of a

6 Maas's analysis of the same phenomenon (Maas 2009) uses the term "complex predicate", Fabri and Borg (2017) refer to this phenomenon as "verbal sequence". I prefer Stolz's more neutral (cf. Stolz 2009: 138) and more poetic terminological choice.

single lexical verb *stenna* "to wait" bearing the brunt of the semantics and three auxiliares: the modal of obligation *ikoll-*, the repetitive *reġa'* and the inchoative *beda*. In syntactic terms, all three are dependents of the lexical verb, thus the dependency graph in 83 above.

And there are two syntactic arguments for analyzing the entire verbal chain in those terms, both having to do verbal modification. The first one can already be observed in (83), but is made much clearer in (84):

(84) *Kultant ma nkunx nista' nibqa' nħares lejh.*
occasionally NEG I was not I can I remain I look at him
'Occasionally, I wasn't able to keep looking at him.'

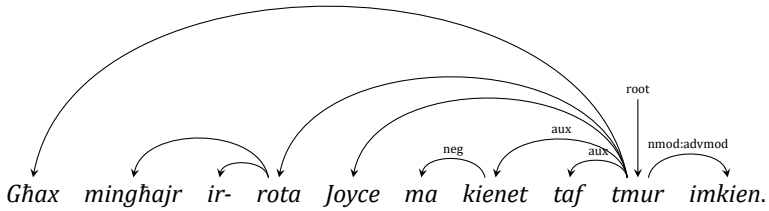


[BCv3: 2008 Loranne Vella Simon Bartolo-Wied Wirdien (Fiddien II)]

In this sentence, the entire verbal chain is negated, but the negation (both the particle and the suffix) is only realized on the first member of the chain. The same applies to the FUT particle in (83): the entire verbal chain refers to the future, but it is only the first verb in the sequence that bears the appropriate marker.

The second argument for the analysis of the verbal chain as a single unit also involves negation, more specifically, the phenomenon of *x*-dropping (Lucas 2014):

(85) *Għax mingħajr ir- rota Joyce ma kienet taf tmur imkien.*
because without DEF bike Joyce NEG she was she knows she goes nowhere.
'Because without a bike, Joyce didn't know how to get anywhere.'



[BCv3: 2011 Trevor Żahra - Qamar Aħdar]

This verbal chain governs the PRON_INDEF (a negative polarity item) *mkien* which in negative sentences is an adverbial meaning "nowhere" and which by its nature as a polarity item triggers *x*-dropping on its verbal governor. As an adverbial of location, however, it can only be a dependent of the main lexical verb *mar*; and yet its effect carries over all the way to the first verb in the chain, *kien*.

All of this speaks in favor of adopting the Many Auxiliaries Theory both in general as well as for the purposes of syntactic annotation. When attempting to do the latter in MUDTv1, however, a number of problems emerged: for one, there was the issue of the large number of elements that can break up the chain (Stolz 2009: 154-157), but that could be overcome by simply applying the UD annotation Principle IV ("UD relations are as flat as possible") and annotate them as siblings to the auxiliaries. Then there was the issue of agreement (Stolz 2009: 147-149), but as this only involves the pseudoverb *kell-/għand-/ikoll-* "to have", it did not present that much of a problem and has in any case been adequately explained (Camilleri 2018). The biggest hurdle to overcome was the sheer number of candidates for auxiliaries: Vanhove's auxiliary candidate list (Vanhove 1993: 153-330) has some 28 items and further candidates kept cropping up, like *ipprova* "to try", *mar* "to go" (and possibly other verbs of motion, cf. Fabri and Borg 2017: 72-76), *xtaq* "to want; to wish", *ħabb* "to love; to wish" and the pseudoverb *għad* "to be still" (see also Camilleri 2016: 358 for further candidates). Stolz (2009: 146) discusses the first two in the context of their participation in Maltese verbal chaining and describes them as "relatively peripheral candidates" noting that "they optionally take subordinators". These observations cut to the heart of the problem: what makes *pprova* and *mar* peripheral candidates? Is it their semantics? If so, how are they any different from, say, *fetaħ* "to open" or *jaf* "to know"? Or is it the fact that they can optionally take a subordinator? The same is true of modals like *ried* "to want", for example; compare the two constructions in (86) and (87), nearly perfectly identical save for the presence of the subordinator/complementizer *li* in (87) (cf. Borg and Azzopardi-Alexander 1997: 32, see also Stolz's ultimately unsuccessful attempt at an analysis of the phenomenon in Stolz 2009: 145).

(86) *Jien ma ridtx nitlaq mil- Libja.*
I NEG I wanted-NEG I exit from-DEF Libya.
'I didn't want to leave Libya.'

[BCv3: l-orizzont.70732]

(87) *Jien ma ridtx li nitlaq mid- dar t' ommi.*
I NEG I wanted-NEG COMP I exit from-DEF house GEN mother-my.
'I didn't want to leave my mother's house.'

[BCv3: 2011 Clemente Zammit - Tieqa fuq it-Triq]

This behavior of some of the verbs on auxiliary candidate list brings us back to the issue of valency frame and the question of "one verb or many verbs" (see section 6.4.4.2.3) above. Consider the verb *ried* which, along with the subject, can take the following types of core dependents: a verb, a verb with a COMP (*li*) and a noun phrase as a direct object, as in (88).

(88) *Mhux għalhekk ridt il- qaħba ċavetta?*
NEG for this you wanted DEF whore key?
'Wasn't it for this that you wanted the fucking key?'

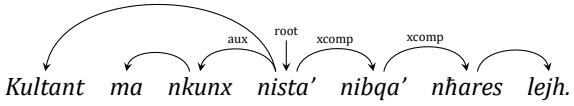
[BCv3: 2007 Mario Azzopardi - Alicia titkellem mill-Imwiet]

What we have here, then, is a proper transitive (bivalent) lexical verb which sometimes (so Vanhove 1993 and Stolz 2009) acts as an auxiliary. But, and here I again focus on the practical issues involved in annotation, how can we tell in a consistent manner? The syntactic criteria (the subordinator/complementizer) are out the window, so are the semantic ones, at least when it comes to the straightforward ones and to establish a set of more detailed conditions or tests for the 30+ lexical verbs on Vanhove's auxiliary candidate list (as expanded above) would require a significant amount of effort. The Many Auxiliaries Theory, while certainly valid (with some necessary elaboration) for descriptive purposes, became impractical when it comes to syntactic annotation.

But what to replace it with? As it turns out, the answer is already contained in Vanhove's auxiliary candidate list: in the previous paragraph, I described it as containing "30+ lexical verbs". This is not entirely correct, especially when viewed from the point of UD: the copular verb *kien* "to be" stands out. In UD (Nivre, Ginter et al. 2014, Nivre, Ginter et al. 2016), copular clauses are treated differently from verbal clauses, in that the copular verb is considered a dependent of the predicate, not the head of the clause (see section 6.4.4.1.3). That leaves *kien* as the odd one out: in UD v1 and thus MUDTv1, it is never (for a given value of "never", see section 6.4.4.4.4 above) the head of a clause and thus it is not treated as a lexical verb. Whenever it functions as a copular verb (i.e. attaches to the predicate in a copular clause), it is labeled cop. And so we are left with the second major type of structures featuring *kien* in which *kien* is governed by a verb (84) and this would the second route to take when deciding what counts as an auxiliary: only KIEN that depends on a VERB will be considered an auxiliary in MUDTv1. This shall henceforth be referred to as One(-ish) Auxiliary Theory: the only aux is *kien*, all the other verbs in verbal chains will attach to the previous one in succession by means of the xcomp relation (see section 6.4.4.4.3 above) with the first non-*kien* verb as the root. For an example, see (84) as converted to One(-ish) Auxiliary Theory in (89) below:

(89) *Kultant ma nkunx nista' nibqa' nħares lejh.*
occasionally NEG I was not I can I remain I look at him

'Occasionally, I wasn't able to keep looking at him.'



[BCv3: 2008 Loranne Vella Simon Bartolo-Wied Wirdien (Fiddien II)]

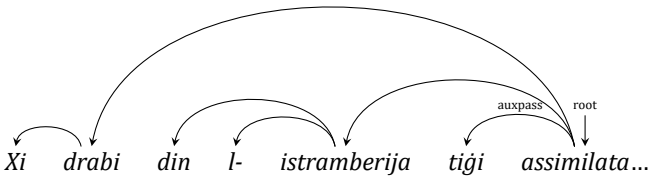
This solution is clean and consistent and if it should turn out to be wrong, it is very easy to correct. It does present a few difficulties, but they can easily be dealt with using UD annotation principles and rules, especially Rule V (see section 6.4.4.2.3 above) which works for both core dependents of what is under Many Auxiliaries Theory referred to as the lexical verb (i.e. the last verb in the chain), as well as for any elements that break the chain (Stolz 2009). As such, the One(-ish) Auxiliary Theory is adopted as the solution for the aux/verbal chain problem in MUDTv1.

In addition to KIEN, three more tokens are tagged as aux in MUDTv1, hence the modifier "-ish" in One(-ish) Auxiliary Theory: the VERB_PSEU *għad*, its negated form *għadx* and the VERB_PSEU *tantx*. The only reason for this is that they don't fit anywhere else: their status as group 2 pseudoverbs (see Chapter 5, section 5.4.1.3.36) is based solely on the fact that they can be negated and the fourth member of this set, *hemm*, receives its own part-of-speech tag and is the root of its own clause.

6.4.4.8.2 Passive auxiliary: auxpass

This label is used exclusively for the verb *ġie* (originally meaning "to come") when it is governed by a PASS_PART to form the dynamic passive (Borg and Azzopardi-Alexander 1997: 214) as in example (90) below.

(90) *Xi drabi din l- istramberija tiġi assimilata...*
some time-PL this.F DEF strangeness she comes assimilated-F...
'Sometimes this strangeness is assimilated...'



[MUDTv1: 50_01N10]

Here, as noted in section 6.4.4.1.2 above, an ambiguity arises between the dynamic passive and the stative passive where the place of *ġie* is taken by KIEN, PRON_PERS or left empty and such constructions can be analyzed simply as featuring a PART_PASS

acting as a predicative adjective. In some cases, this ambiguity can be resolved at the level of predicate arguments, i.e. by marking the overt subject of the clause as nsubj or nsubjpass and/or by marking the agent as nmod:agent. In others, the ambiguity is only resolvable based on the analysis of the valency of the root PART_PASS (as per section 6.4.4.1.2). Due to this ambiguity, KIEN and PRON_PERS in the stative passive are always labeled cop and not auxpass, if only for the sake of simplicity and consistency.

And finally, as Vanhove (1993: 324) notes, there is one more type of a passive structure, that featuring the verb *safa* + PART_PASS. Vanhove describes this construction as extremely rare (and is able to cite only three examples) and corpus data confirms this: there are fewer than 5000 examples (27 per million) in *BCv3* and none in the texts selected for MUDTv1. Should this structure be encountered at any point in the future development of MUTDv1, *safa* will also be labeled as auxpass.

6.4.4.8.3 Copula: cop

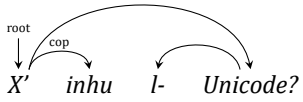
This relation is used for copulas in copular clauses, i.e.:

i. PRON_PERS in type (ii) copular clauses;
ii. *qiegħed* in all its forms in type (iii) copular clauses; and
iii. KIEN in type (iv) copular clauses.

For examples, see the discussion in section 6.4.4.1.3 above.

In addition to these, the cop relation is also used for PRON_INT *inhu*, *inhi* and *inhuma* which serve as copulas in interrogative copular clauses of identity (91).

(91) *X'* *inhu* *l-* *Unicode?*
what COP-INT DEF Unicode?
'What is Unicode?'

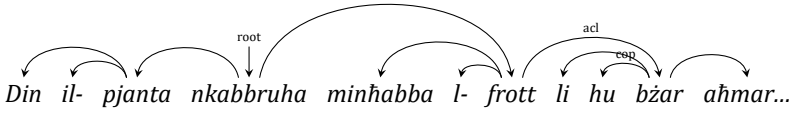


[MUDTv1: 57_04N11]

And finally, the cop relation is also employed for the solitary PRON_PERS dependent of copular acl clauses (92).

(92) *Din* *il-* *pjanta* *nkabbruha* *minħabba* *l-* *frott* *li* *hu* *bżar* *aħmar...*
this.F DEF plant we grow-ACC.3S.F because DEF fruit COMP he pepper red...

'We grow this plant because of its fruit which is a red pepper...'



[MUDTv1: 56_03N11]

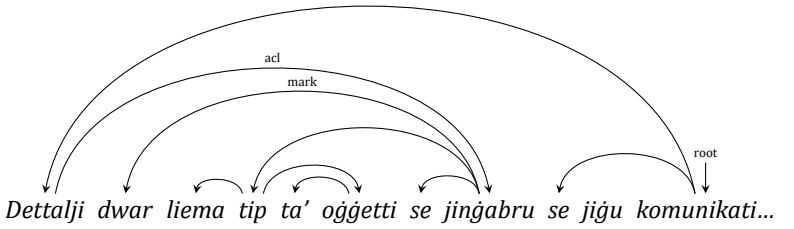
In most clauses of this type, including main clauses, such a PRON_PERS would be interpreted as a nsubj. In copular acl clauses, however, the same position can also be occupied by the copular verb KIEN labeled cop, and so the same label is applied here. The other option, that the PRON_PERS is a nsubj, can be ruled out outright in light of the analysis of relative clauses in Borg and Azzopardi-Alexander (1997: 37).

6.4.4.8.4 Subordinators and complementizers: mark

The mark relation is used for function words introducing subordinate clauses of any type. In MUDTv1, this includes:

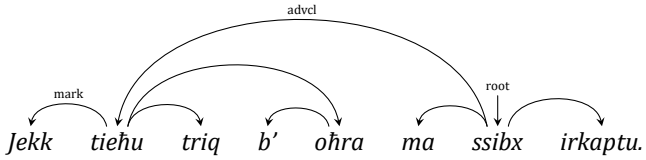
i. tokens tagged COMP in acl, xcomp and ccomp (92);
ii. tokens tagged PREP and GEN in acl (93);
iii. tokens tagged CONJ_SUB in acl, advcl (94), xcomp and ccomp;
iv. tokens tagged PREP with a COMP dependent in a mwe relation in advcl (124);
v. tokens tagged ADV (mostly *hekk* or *aktar*) with a dependent in a mwe relation in advcl; and
vi. tokens tagged PRON_PERS with a *u* "and" dependent in a mwe relation in advcl.

(93) *Dettalji dwar liema tip ta' oġġetti se jinġabru se*
detail-PL about which type GEN object-PL FUT they are collected FUT
jiġu komunikati aktar tard.
they come communicated-PL more late
'Details about which type of objects will be collected will be communicated later.'



[MUDTv1: 08_08J01]

(94) *Jekk tieħu triq b' oħra ma ssibx irkaptu.*
if you take road with other NEG you find-NEG gear.
'If you take one road after another, you will not find the gear.'



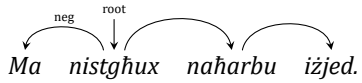
[MUDTv1: 49b_04F09]

6.4.4.8.5 Negation: neg

This relation is used for negators, which include:

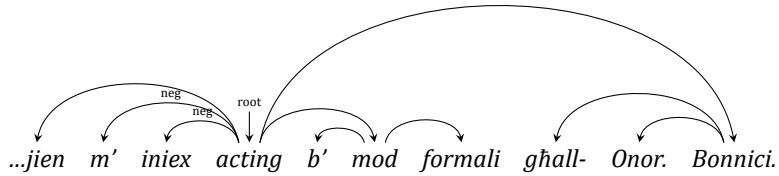
i. tokens tagged NEG, i.e. the particle *ma* and its variant *m'* (95);
ii. tokens tagged PRON_PERS_NEG (96); and
iii. the particle *la* in the *la ... lanqas* construction (97).

(95) *Ma nistgħux naħarbu iżjed.*
NEG we can-NEG we run more.
'We cannot run anymore.'



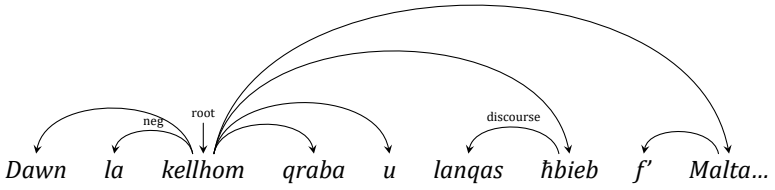
[MUDTv1: 47_01F09]

(96) *...jien m' iniex acting b' mod formali għall- Onor.*
...I NEG I-NEG acting with manner formal on-DEF Right Honorable
Bonnici.
Bonnici.
'...I am not acting in a formal manner in Right Honorable Bonnici's place.'



[MUDTv1: 38_02P06]

(97) *Dawn la kellhom qraba u lanqas ħbieb f' Malta...*
these.M NEG they had relative.PL and nor friend.PL in Malta...
'These people had neither relatives, nor friends in Malta...'

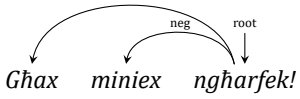


[MUDTv1: 51_02N10]

Two remarks concerning PRON_PERS_NEG as neg: first, there exists a three-way variation in the spellling of PRON_PERS_NEG in BCv3 when it comes to whether the negative particle *ma/m'* is split off or not; e.g. for 3rd person singular masculine, the options are *mhuwiex* (one token, 20,099 occurrences in *BCv3*), *m'huwiex* (two tokens, 15,246 occurrences in *BCv3*) and *ma huwiex* (two tokens, 203 occurrences in *BCv3*). Whenever a two-token form is used, this results in two neg dependents on a single governor, as in (96). This is not ideal, but not especially problematic either, and so this is the solution adopted in MUDTv1.

Secondly, in light of its use and the PRON_PERS / PRON_PERS_NEG pair where the former functions as a copula, it may seem more proper to use a separate relation for PRON_PERS_NEG, something along the lines of cop:neg. This, however, would not be appropriate: PRON_PERS_NEG negates not only copular predicates, but is also used as a verbal negator, whether in conjunction with verbal particles as in (99), which may be its original function relating to the original function of the particles as participles, or negating a verb directly (98) (see Al-Sayyed and Wilmsen 2017 for a detailed analysis). A simple and consistent solution to annotate all occurrences of PRON_PERS_NEG (including those in non-expletive subjectless clauses, see section 6.4.4.1.6) as neg therefore seemed like the best choice.

(98) *Għax miniex ngħarfek!*
because I-NEG I know-ACC.2SG!
'Because I don't know you!'



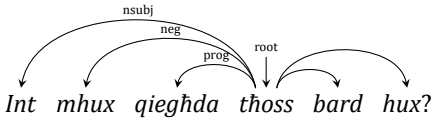
[BCv3: 57]

6.4.4.8.6 Verbal particles: part

This relation covers the verbal particles FUT (*se* and its variants, *ħa* and *għad* in the appropriate context, see Chapter 5, section 5.4.1.3.10) and PROG (i.e. *qed* and its variants, see Chapter 5, section 5.4.1.3.26). Some UD v1 treebanks, like the Bulgarian one, subsume these words under the aux label. In the light of the complexities of the category aux detailed above, I found it preferable to establish a separate category for these particles.

This decision was also motivated by syntactic and morphological considerations: as for *qed* and its variants *qiegħed*, *qiegħda* and *qegħdin*, the latter three are morphologically present active participles. As such, they have three functions: in the first, they are copulas, in which case their part of speech is PART_ACT and they are labeled as cop (see sections 6.4.4.1.3 and 6.4.4.8.3 above). In the second, they are existential predicates and they are the root in their respective clause (see section 6.4.4.1.5 above). And finally, their third function is to modify a verb to indicate the ongoing nature of the action or process, the same way *qed* does (Vanhove 1993: 113-134). This behavior is very similar to that of KIEN and so the label of aux might very well be extended to those three words. There is, however, one significant exception here: unlike KIEN (and the other auxiliary verb in MUDTv1, *għad*), active participles cannot bear markers of negation and so whenever a verbal catena *qiegħed*, *qiegħda* and *qegħdin* appear in needs to be negated, PRON_PERS_NEG must be used:

(99) *Int mhux qiegħda tħoss bard hux?*
you NEG PROG-F you feel cold right
'You are not feeling cold, right?'



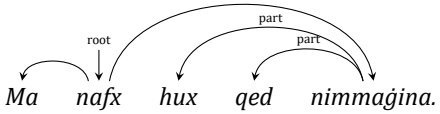
[BCv3: dance-dance-dance]

This, the relatively straightforward division of labor and the general non-finite nature of all forms of *qiegħed*, plus all the considerations cited in the section on aux confirm the choice of establishing part as a separate relation for these verbal modifiers.

The same analysis was also applied to the particles *se* and *ser* and the present active participle *sejjer* it is ultimately derived from. In this case, however, *sejjer* (together with its feminine form *sejra* and plural *sejrin*) is also a little more syntactically flexible than *qiegħed* (e.g. it can be used as an attributive adjective), and unlike *qiegħed*, it can function as a proper predicate in terms of UD v1. It is therefore treated as a PART_ACT and a head of its own clause whenever it does not directly modify a verb. When it does, it is – together with all its forms and *se* and *ser* – annotated as part.

And finally, the part relation is also used for PRON_PERS with the interrogative suffix *-x* attached when governed by a verb or a pseudoverb, as *hux* in (100):

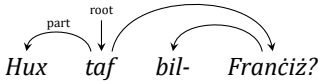
(100) *Ma nafx hux qed nimmaġina.*
NEG I know-NEG he-INTR PROG I imagine.
'I don't know if I'm imagining things.'



[BCv3: 2012 Ivan Buġeja - Ġimgħa Sibt u Ħadd]

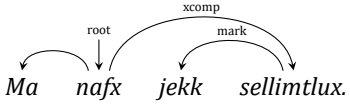
Such PRON_PERS typically serve as copulas, but that can hardly be the case here. When governed by a verb, they are either interrogatives (101) or subordinators (100). In their latter use, they could be analyzed as mark (see section 6.4.4.8.4 above), especially since they fulfill the same function as (100) and are in complementary distribution with the subordinator *jekk* (102).

(101) *Hux taf bil- Franċiż?*
he-INTR you know with-FEF French?
'Do you know French?'



[BCv3: it-torca.42728]

(102) *Ma nafx jekk sellimtlux.*
NEG I know-NEG if I greeted him-INTR.
'I don't know if I greeted him.'



[BCv3: 2010 Trevor Zahra - Fuklar qadim u bnadar imċarta]

The fact that in constructions like (102) the verb features the interrogative suffix *-x* suggests an alternative analysis: in clauses like (100), the PRON_PERS actually serves as a skeleton to hang the interrogative suffix on. That such constructions featuring KIEN with the interrogative suffix and *hux* et al. can also be used in main clauses (101) only goes to confirm the function of PRON_PERS + interrogative *-x* as an alternative for the

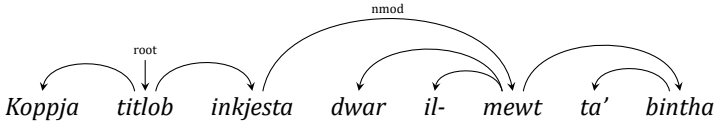
verbal interrogative suffix *-x*. As such, they behave in much the same way as neg PRON_PERS_NEG does when acting as a verbal negator, in that it suppresses the negative suffix *-x* on the verb (98). In other words, both the interrogative suffix *-x* and its homophone/homograph negative suffix can either be attached to the verb or they can be borne by a PRON_PERS, but not both. In the latter case, the PRON_PERS with the negative suffix attached is tagged as PRON_PERS_NEG and in the neg relation to its governor. In case of PRON_PERS with the interrogative suffix *-x*, there is no special part-of-speech tag (though perhaps there should be) and for the relation to its governor, I decided to use the part rather than aux.

6.4.4.9 Nominal dependents: Nominals

6.4.4.9.1 Nominal modifier: nmod

As noted in section 6.4.4.2.3, the nmod relation is used for all noun-headed dependents of anything but VERB, VERB_PSEU, PART_PASS, PART_ACT and KIEN that do not qualify as nmod:poss, conj, appos or nmod:advmod. These are typically modifiers, as the prepositional phrase in (103), but the semantics does not play a role here, only the two syntactic criteria are used in determining what a nmod is.

(103) *Koppja titlob inkjesta dwar il- mewt ta' bintha*
couple she requests inquiry about DEF death GEN daughter-her
'Couple requests inquiry into daughter's death'

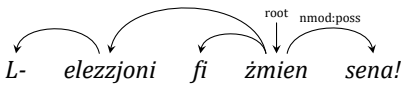


[MUDTv1: 05_05J01]

6.4.4.9.2 Possessive nominal modifier: nmod:poss

As noted in section 6.4.4.2.3, the nmod:poss relation was split off nmod for possessive constructions, whether of the analytical type mediated by GEN and GEN_DEF (Borg and Azzopardi-Alexander 1997: 76) or the construct state (Borg and Azzopardi-Alexander 1997: 71). In these constructions, nmod:poss marks the possessed elements introduced by GEN or GEN_DEF (105) or the *nomen rectum* (104).

(104) *L- elezzjoni fi żmien sena!*
DEF election in time year!
'The election is in a year's time!'

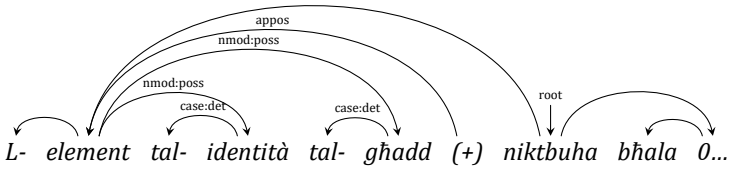


[MUDTv1: 17_04J02]

6.4.4.9.3 Apposition: appos

UD v1 annotation guidelines define apposition as "a nominal immediately following the first noun that serves to define or modify that noun" (Nivre, Ginter et al. 2014). UD v2 guidelines (Nivre, Ginter et al. 2016) update this to "a nominal phrase that follows the head of another nominal phrase and stands in a co-reference or other equivalence relation to it" and the UD v2 validation rules[7] mandate that appos always be left-headed, without the condition of immediacy. This condition often cannot be satisfied in Maltese, as various modifiers routinely follow the noun (105), and so the UD v2 rules were used for MUDTv1.

(105) *L- element tal- identità tal- għadd (+) niktbuha bħala 0...*
DEF element GEN-DEF identity GEN-DEF addition (+) we write her as 0...
'The identity element for addition (+) is written as 0...'



[MUDTv1: 55_02N11]

The appos relation is also used for noun phrases introduced by the PREP_PRON *fosthom* "among them" which defines a set by naming at least one member.

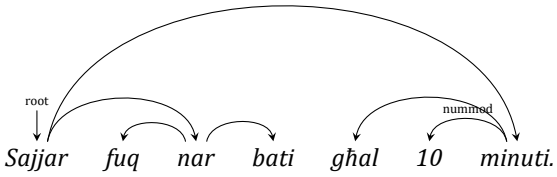
6.4.4.9.4 Numeral modifier: nummod

This relation is used for all numerals (i.e. tokens tagged NUM_CRD and X_DIG) which modify a noun or an equivalent word class (e.g. X_PUN or X_ABV). This includes the year (106), based on the treatment of such constructions in other UD v1 treebanks.

(106) *Sajjar fuq nar bati għal 10 minuti.*
cook on fire low on 10 minute-PL.

7 bit.ly/2EWIREQ (last consulted on February 28th 2018)

'Cook on low flame for 10 minutes'



[MUDTv1: 54_01N11]

This does not apply to tokens tagged NUM_ORD which are in an amod relation to their governor, nor to tokens tagged NUM_FRC which are in the construct state with the noun they count.

6.4.4.10 Nominal dependents: Clauses

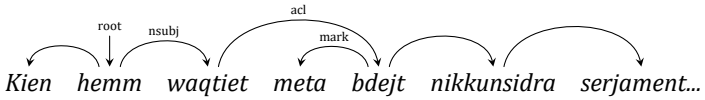
6.4.4.10.1 Adjectival clause: acl

Adjectival clauses are verbal, copular or existential clauses of any type that modify a noun (i.e. NOUN and equivalent word classes such as NUM.*, QUAN, X_ABV or X_PUN) or a pronoun. These come in two major types:

I. Adjective clauses (Borg and Azzopardi-Alexander 1997: 35-37), and
II. headless relative clauses (Borg and Azzopardi-Alexander 1997: 37).

Adjective clauses proper are typically introduced by the complementizer/subordinator *li* (38), but may also be introduced by CONJ_SUB like *biex* or *meta* (107), GEN (*ta'*) and PREP such as *dwar*.

(107) *Kien hemm waqtiet meta bdejt nikkunsidra serjament li nfittex xogħol ieħor.*
he was EXIST time-PL when I began I consider seriously COMP I find work other.
'There were times when I began to seriously consider finding a different job.'



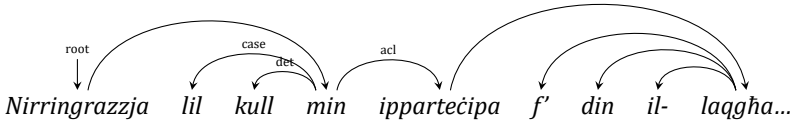
[MUDTv1: 49_03F09]

Additionally, there is also a subtype where the acl consists of a verb without a subordinator (Borg and Azzopardi-Alexander 1997: 35, 60).

Headless relative clauses modify the indefinite pronouns *min* "he/she who" and *xi* "it that". In MUDTv1, the PRON_INDEF in question are treated as the head of the phrase,

on the logic that they can take case markers and determiners, as with *lil* and *kull* in (108):

(108) *Nirringrazzja lil kull min ipparteċipa f' din il- laqgħa...*
I thank ACC all he who he participated in this DEF meeting...
'I thank everyone who participated in this meeting...'



[MUDTv1: 30_01P05]

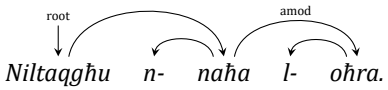
And finally, participles (whether PART_PASS or PART_ACT) which follow a nominal and depend on it without a complementizer are treated as acl only if they contain a dependent: the logic behind this is that while participles can assume the same attributive role as adjectives, unlike those adjectives that follow a noun, they can also be roots of verbal or copular clauses and thus can take arguments like nmod:agent or nmod:advmod.

6.4.4.11 Nominal dependents: Modifier words

6.4.4.11.1 Adjectival modifier: amod

This relation is used for tokens tagged as ADJ modifying a noun or a pronoun regardless of whether they follow (109) or precede their head, PART_ACT which invariably follow their head (110) and PART_PASS when they precede their head or when they follow it and they do not take any arguments (111).

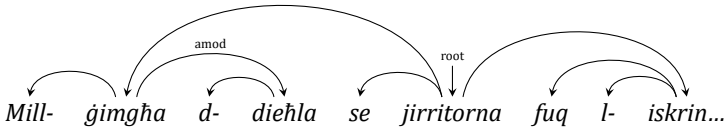
(109) *Niltaqgħu n- naħa l- oħra.*
we meet DEF side DEF other-F
'We meet on the other side.'



[MUDTv1: 49b_04F09]

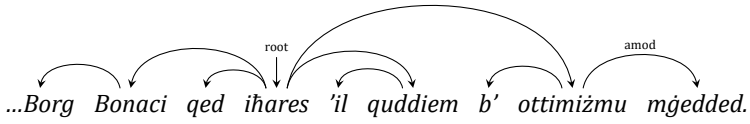
(110) *Mill- ġimgħa d- dieħla se jirritorna fuq l- iskrin...*
from-DEF week DEF coming FUT he returns on DEF screen...

'Starting next week, he returns to the screen...'



[MUDTv1: 21_01J03]

(111) *...Borg Bonaci qed iħares 'il quddiem b' ottimiżmu mġedded.*
...Borg Bonaci PROG he looks to front with optimism renewed.
'... Borg Bonici looks ahead with renewed optimism.'



[MUDTv1: 21_01J03]

Along with these, the label amod is used for ordinal numerals (NUM_ORD) and the quantifier (QUAN) *ebda* "no, none" on the logic that like adjectives, they can take a definite article as a dependent.

6.4.4.11.2 Adverbial modifier: advmod

This relation is used for tokens tagged ADV that modify nominals and their dependents, mostly those in amod relation to them. As ADV that modify verbs are also labeled advmod, all ADV are in effect attached to their governor through the advmod relation.

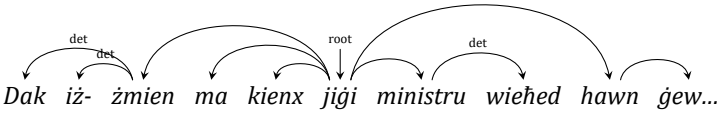
6.4.4.12 Nominal dependents: Function words

6.4.4.12.1 Determiner: det

This relation is used for determiners which in MUDTv1 come in the following types:

I. The definite article (DEF),
II. modifers that appear in complement to the definite articles, i.e. quantifiers (QUAN), the numeral wieħed (NUM_WHD) when modifying a noun or equivalent, and interrogative pronouns (PRON_INT) *x'*, *liema* and *kemm* when modifying a noun or equivalent, and
III. demonstrative pronouns (PRON_DEM), including their forms with fused definite article (PRON_DEM_DEF).

(112) *Dak iż- żmien ma kienx jiġi ministru wieħed hawn ġew...*
that DEF time NEG he was-NEG he comes minister one here inside...
'Back then, not a single minister came here...'



[MUDTv1: 38_02P06]

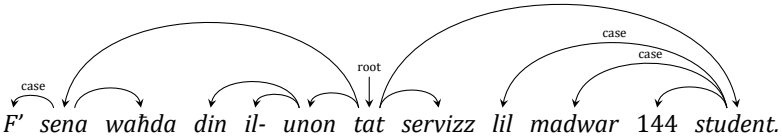
As evident from the comparison to part-of-speech criteria, there is some mismatch here between word classes and UD v1 labels: for example, *ebda* "no, none" is tagged as QUAN, but treated as amod, not det, for the purposes of syntactic annotation. The logic behind it is that syntactically, it behaves as an adjective, since it takes a definite article itself. This mismatch is the result of the fact that for part-of-speech tagging, syntactic criteria are only take into account after semantic and morphological considerations have been exhausted (see Chapter 6, section 6.4.1.2), whereas for syntactic annotation, they obviously come first.

6.4.4.12.2 Case: case

In general, the case relation is used for all function words that mark case relations in one way or another. In MUTDv1, this includes:

i. The genitive marker *ta'* (GEN), including its forms with fused pronouns (GEN_-PRON),
ii. the oblique marker *lil* (LIL), and
iii. prepositions (PREP).

(113) *F' sena waħda din il- unon tat servizz lil madwar 144 student.*
in year one-F this.F DEF union she gave service DAT about 144 student.
'In one year, this union provided service to about 144 students.'

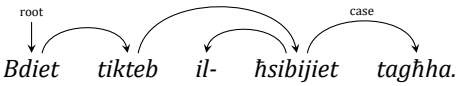


[BCv3: 20091121_166d_par]

The genitive marker with fused pronouns (GEN_PRON) stands out from among the other members of the category, since their respective forms with fused pronouns (LIL_-

PRON and PREP_PRON) are treated as pronouns. This treatment of GEN_PRON is motivated by two considerations: first, unlike LIL_PRON and PREP_PRON which by themselves can be heads of phrases, GEN_PRON is – discounting ellipsis[8] – always a dependent, never the head. Secondly, here I took cues from Hebrew: in the Hebrew UD 2.1 treebank (Nivre, Agić et al. 2017), the Hebrew adposition *šel* with attached pronouns, which is by and large functionally identical to Maltese GEN_PRON, is also treated as case, albeit a special category (case:gen).

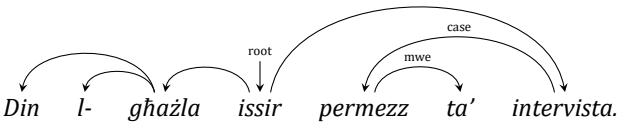
(114) *Bdiet tikteb il- ħsibijiet tagħha.*
she.began she.writes DEF thought-PL her
'She began to write down her thoughts.'



[BCv3: 263]

This relation is also used for those adjectives or nouns that essentially function as prepositions (cf. the analysis in Stolz 2017), but connect to their governor by means of the genitive particle *ta'*. These are primarily *qrib* and *permezz*, where the former alternates between connecting to its governor directly (i.e. *qrib* + NOUN) and connecting to its governor using *ta'* (i.e. *qrib ta*' NOUN). In both cases, *qrib* is tagged as PREP as opposed to *permezz*, which is tagged as NOUN throughout (see chapter 6, section 6.4.1.3.25). In MUDTv1, howevery, they are both considered direct dependents of their governor in the case relation to it. The entire construction with the genitive particle *ta'* is then analyzed as a multi-word expression by analogy with mark composed of PREP + COMP (see above) and *ta'* is thus marked as mwe dependent on *qrib* or *permezz*.

(115) *Din l- għażla issir permezz ta' intervista.*
this.F DEF choice it.happens means GEN interview.
'This choice happens by means of an interview.'



[BCv3: 20101108_277d_par]

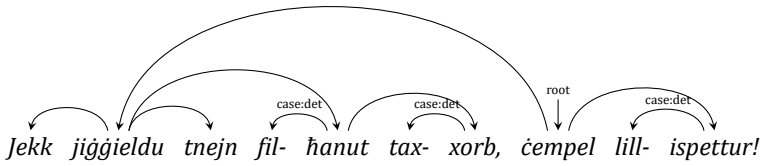
8 As in *Idejh imċappsa bid- demm, mhux minn* ***tiegħu/GEN_PRON****, iżda minn ta' sid il- ħatar.* [MUDTv1: 47_01F09] "His hands stained with blood, not from **his**, but from the owner of the stick". Here it is assumed that the actual head of the GEN_PRON (*demm* "blood") has been ellided and the GEN_PRON is thus, in line with UD v1 guidelines (Nivre, Ginter et al. 2014), promoted to the head.

6.4.4.12.3 Case fused with determiner: case:det

This relation is used for all functions words that mark case relations in one way or another and are fused with the definite article. In MUDTv1, this includes:

i. The genitive marker *ta'* with fused definite article (GEN_DEF),
ii. the oblique marker *lil* with fused definite article (LIL_DEF), and
iii. prepositions with fused definite article (PREP_DEF).

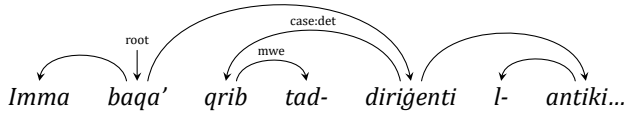
(116) *Jekk jiġġieldu tnejn fil- ħanut tax- xorb, ċempel lill- ispettur!*
if they fight two in-DEF store GEN-DEF drink, call ACC-DEF inspector!
'When two guys fight in a liquor store, call the inspector!'



[BCv3: 1986 Oliver Friggieri - Fil-Parlament ma Jikbrux Fjuri]

This relation is also used for multiword prepositions headed by *qrib* and *permezz* (see above) when their *ta'* component is fused with the definite article, as in (117).

(117) *Imma baqa' qrib tad- diriġenti l- antiki...*
but he remained close GEN-DEF mover-PL DEF ancient-PL...
'But he remained close to the old movers...'



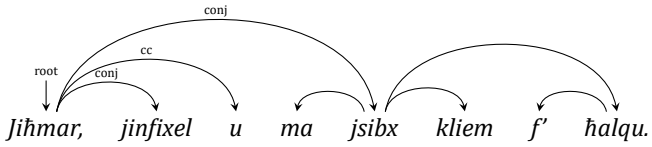
[MUDTv1: 20_07J02]

6.4.4.13 Coordination

6.4.4.13.1 Conjunct: conj

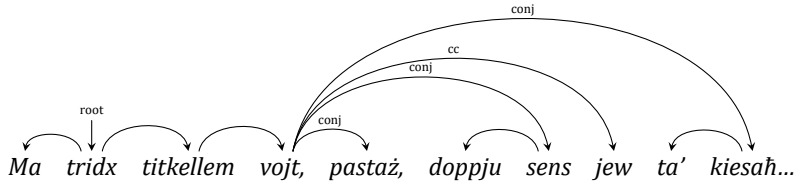
In UD, coordination is treated as an asymmetric relationship, in that the first member is the head and all coordinates, as well as the coordinating conjunction, attach to it. The conj relation is used for all types of coordination, be their between clauses (118), phrases or words or any combination thereof (119).

(118) *Jiħmar, jinfixel u ma jsibx kliem f' ħalqu.*
he reddens, he gets confused and NEG he finds-NEG word.PL in throat-his.
'He blushes, gets confused and stutters.'



[MUDTv1: 40_02F08]

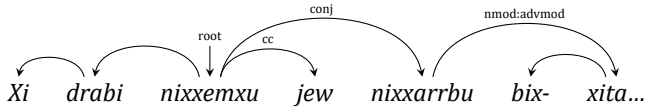
(119) *Ma tridx titkellem vojt, pastaż, doppju sens jew ta'*
NEG you want-NEG you speak empty, vulgar, double meaning or GEN
kiesaħ...
cold...
'You wouldn't speak in a shallow or vulgar manner, misleadingly or coldly...'



[MUDTv1: 21_01J03]

In coordinations (especially those involving verbs) which have further dependents, those dependents attach to the head of the coordination by default, unless the valency of the word or semantic considerations make it clear that they are dependents of the respective conjunct (120).

(120) *Xi drabi nixxemxu jew nixxarrbu bix- xita...*
some time-PL we sunbathe or we get soaked with-DEF rain...
'Some days we bask in the sun or get soaked in the rain...'



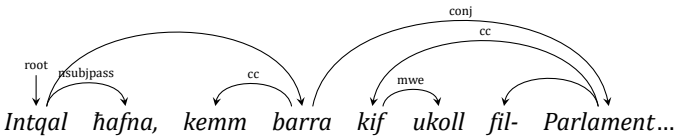
[MUDTv1: 42_04F09]

6.4.4.13.2 Coordinating conjunction: cc

The cc relation is used for coordinating conjunctions, i.e. the tokens tagged CONJ_CORD. Since in UD v1, the first conjunct is the head of the conjunction, the CONJ_CORD attaches to the first conjunct (Nivre, Ginter et al. 2014), as *jew* "or" in (119). This is inconsistent with the analysis of similar functions words (see mark) and was changed in UD v2 so that a CONJ_CORD attaches to the member of coordination they belong to (Nivre, Ginter et al. 2016), but for MUDTv1, the UD v1 guidelines apply.

In the *kemm ... kif ukoll* construction and multiple-part coordinations of the *kemm ... kif ... kif ukoll* type, both (or all) constituent parts are tagged as cc. Contrary to the usual behavior of cc in UD v1, each attaches to its respective member of the coordination, e.g. *kemm* attaches to the first one, *kif ukoll* attaches to the second one as in example (refcc2); the latter is true even in situations where *kif ukoll* appears on its own. In terms of internal structure, *ukoll* is dominated by *kif* in the mwe relation; this is by analogy with PREP + COMP constructions like *wara li* (see below).

(121) *Intqal ħafna, kemm barra kif ukoll fil- Parlament...*
it is said much, as outside as well as in-DEF Parliament...
'Much is said, outside as well as in the Parliament...'



[BCv3: ilgensillum.2011-Jannar-18.4446]

6.4.4.14 Multi-word expressions

6.4.4.14.1 Compound: compound

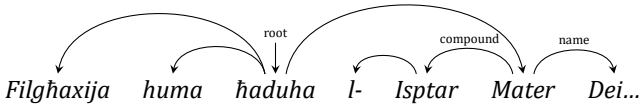
In UD v1, the labor on multi-word expressions is divided between three relations. The first of them, compound, is used for content words, and in MUDTv1, this is primarily applied to two types of constructions:

i. Appositional compounds of the NOUN_{set} + NOUN_{member} type
ii. Light Verb Constructions (LVCs)

In appositional compounds, the compound relation is used for the set designation as with *l-isptar* in (122), the logic being that it merely provides additional information and can thus be eliminated without changes to the semantic content of the phrase.

(122) *Filgħaxija huma ħaduha l- Isptar Mater Dei...*
in the evening they they took her DEF hospital Mater Dei

'In the evening, they took her to the Mater Dei hospital...'

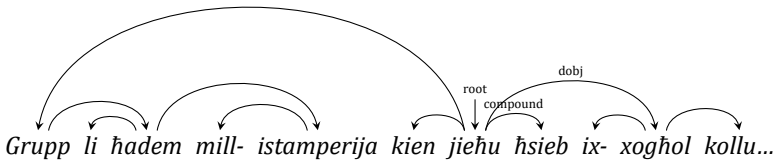


[MUDTv1: 05_05J01]

Dependents which are governed by the compound as a whole (LIL, LIL_DEF, PRON and PRON_DEF) attach to the subordinate word. This goes against UD Principle IV ("UD relations are as flat as possible") and is somewhat clunky, but since it does not cause any obvious problems, it will remain the preferred solution.

The other type of multi-word expression for which the compound relation is used in MUDTv1 are the so-called Light Verb Constructions (LVCs). These are verb-headed multi-word expressions with a noun dependent displaying a number of idiosyncratic syntactic and semantic properties, chief among them the empty semantics of the verb; in other words, the verb serves only to express grammatical meanings such as tense, but does not add any semantic content to the noun (Savary et al. 2018). The analysis of Maltese LVCs in Čéplö and van der Plas 2017 based on the annotation guidelines of the *PARSEME shared task on automatic identification of verbal MWEs - edition 1.0 (2017)*[9] has shown that the issue of LVC identification in Maltese is somewhat complex and requires further work. In light of this, the decision was taken to only apply the compound label to a number of clear-cut cases, such as *ħa ħsieb* "to take care" (123), *ta sehem* "to take part", *għand-/kell-/ikoll- bżonn* "to need" and *ħa pjaċir* "to take pleasure".

(123) *Grupp li ħadem mill- istamperija kien jieħu*
group COMP he works from-DEF printing house he was he takes
ħsieb ix- xogħol kollu...
thought.PL DEF work all of him...
'A group that worked out of a printing house used to take care of the whole job...'



[MUDTv1: 20_07J02]

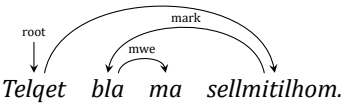
9 bit.ly/2CkIHWd, last consulted on February 28th 2018

In addition to these two functions, the compound relation is also used for absolute reduplication of content words; there are only two instances of this in MUDTv1.

6.4.4.14.2 Multi-word expression: mwe

This relation is used in multi-word expressions that serve as function words. In MUDTv1, these are predominantly case markers such as *permezz ta'* (see section 6.4.4.12.2) or subordinators composed of PREP + COMP such as *wara li* "after" or *bla ma* "without X-ing". The mwe relation is used for the second part of such multi-word expressions while the first attaches to its head and is marked with the syntactic function it performs (124).

(124) *Telqet bla ma sellmitilhom.*
she left without COMP she greeted-DAT.3PL
'She left without saying goodbye to them.'



[MUDTv1: 41_03F08]

In addition, this relation is also used for function words that go together and cannot be assigned any other relation, e.g. combinations of two tokens of the same word class (such as the two CONJ_SUB *li kieku* "if only", the two PREP *flimkien ma'* "together with" or the two FOC *lanqas biss* "not even"), the adverb *x' aktarx* "rather" and *kif ukoll* "as well" in its function as a coordinator. And finally, mwe is also used for the combination of PRON_PERS and CONJ_CORD *u* when they combine to act as a subordinator in the *ḥāl* construction (Yoda 2017).

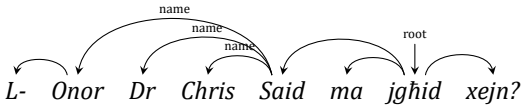
6.4.4.14.3 Name: name

This relation is used for multi-word expressions that consist of proper nouns, i.e. names of people or entities. This only applies to those names that are not analyzable in terms of Maltese syntax; phrases that are (such as the name of the newspaper *L-Orizzont* "DEF-horizon") are annotated according to their constituent parts.

For the former group, the surname (or the last token tagged NOUN_PROP in the sequence in question) is considered the head and all the other names, including initials and titles tagged as X_ABV (such as *Dr.*, *Mrs* or *L-Onor.* "the Right Honourable"), are direct dependents of it (125). For entities (company names etc.), the first word in the name is considered the head, as the name of the *Mater Dei* hospital in (122).

(125) *L- Onor Dr Chris Said ma jgħid xejn?*
DEF RH Dr. Chris Said NEG he says nothing?

'The RH Dr. Chris Said says nothing?'



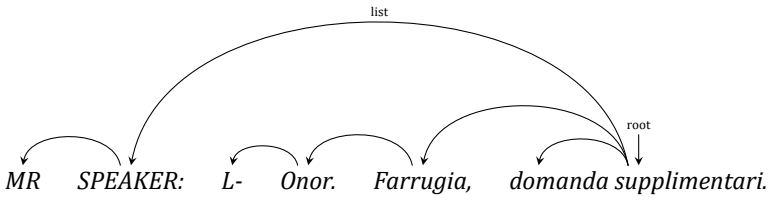
[BCv3: l-orizzont.98220]

6.4.4.15 Loose relations

6.4.4.15.1 List: list

In MUDTv1, this relation is used for flat dependencies that are not clauses (for which parataxis is reserved) or anything else. These include, but are not limited to, structures like list delimiters, *am*/*pm* modifications to expressions of time and mathematical expressions. list is also used in MUDTv1 parliamentary texts where it is employed for the annotation of speaker identification, as in (126).

(126) *MR SPEAKER: L- Onor. Farrugia, domanda supplimentari.*
Mr. Speaker: DEF RH Farrugia, additional question.
'Mr. Speaker: The RH Farrugia, additional question.'

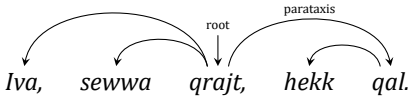


[MUDTv1: 38_02P06]

6.4.4.15.2 Parataxis: parataxis

This relation is used in two scenarios: first, it is used for clauses that are in no clear coordinating or subordinating relation to another (what UD v1 guidelines call "run-on sentences") as in (127). In such cases, the first clause is considered the main clause.

(127) *Iva, sewwa qrajt, hekk qal.*
yes, correctly you read, thus he said.
'Yes, you read that correctly, that's what he said.'



[MUDTv1: 14_01J02]

It should be noted that the mere absence of a conjunction does not a parataxis make, as there are many coordinations that do not use conjunctions. In MUDTv1, parataxis was applied only once all available options were exhausted.

Secondly, this relation is used in reported speech sentences for the clause which contains the actual speech verb, but only if such clause follows the clause containing the reported speech. If such clause precedes the reported speech, the ccomp relation is used instead. The logic behind that only sentences with the speech verb clause at the beginning can be embedded (see the respective entry in Nivre, Ginter et al. 2014).

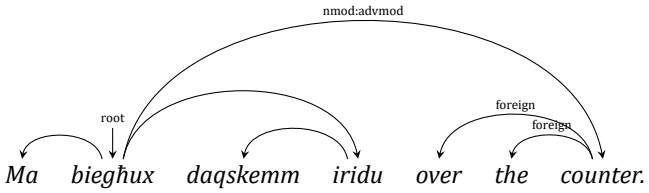
Contrary to the usage in both UD v1 and UD v2, parataxis is only used for clauses in MUDTv1. For flat relations such as news article bylines where, so UD v2 guidelines (Nivre, Ginter et al. 2016), "[t]here does not seem to be a better relation to use", list is used instead.

6.4.4.16 Special relations

6.4.4.16.1 Foreign words: foreign

As noted in Chapter 5, section 5.4.1.3.40, sequences of tokens in English (and, by extension, other languages) that display their native syntax (e.g. they contain a preposition or a verbal argument), are tagged X_ENG or X_FOR. In MUDTv1, such sequences are analyzed in accordance with their native syntax and the head of the phrase is incorporated into the structure of the Maltese sentence in which they feature as much as possible (128).

(128) *Ma biegħux daqskemm iridu over the counter.*
NEG they sold as much as they want over the counter.
'They didn't sell as much as they wanted over the counter.'

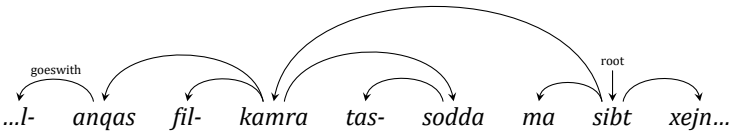


[MUDTv1: 17_04J02]

6.4.4.16.2 Goes with: goeswith

According to the UD v1 annotation guidelines, this relation should be used to link "two parts of a word that are separated in text that is not well edited" (Nivre, Ginter et al. 2014). This is also its primary function in MUDTv1, except in most of the 11 cases where it is used, the underlying cause is not so much sloppy editing, but tokenizer issues (such as a decimal separated from its whole number) or misspellings (as the *l-anqas* "not even" in (129) where the correct form is *lanqas*).

(129) *...l- anqas fil- kamra tas- sodda ma sibt xejn...*
...DEF not even PREP-DEF room GEN-DEF bed NEG I found nothing...
'... not even in the bedroom did I find anything ...'



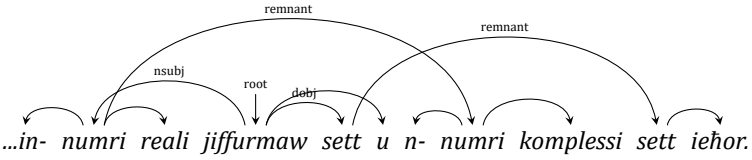
[MUDTv1: 44_06F09]

In addition to that, goeswith is also used for particles in English words that are otherwise integrated into the Maltese sentence, such as phrasal verbs.

6.4.4.16.3 Remnant in ellipsis: remnant

This relation is used "provide a satisfactory treatment of ellipsis" (Nivre, Ginter et al. 2014). The priority for the analysis of ellipsis in UD v1 is to maintain the correct clause and argument structure; in ellipsis, the elements that are preserved are treated as pairs and connected through the remnant relation where the first one serves as the head (130).

(130) *...in- numri reali jiffurmaw sett u n- numri komplessi sett ieħor.*
...DEF number-PL real-PL they form set and DEF number-PL complex-PL set other.
'...real numbers form a set and complex numbers another set.'



[MUDTv1: 55_02N11]

In this way, it is quite easy to discern the original structure of both original clauses, as well as the relationship between them. This is at the expense of creating non-projective dependencies, but the trade-off is acceptable.

6.4.4.16.4 Overridden disfluency: reparandum

This relation is used to "indicate disfluencies overridden in a speech repair" (Nivre, Ginter et al. 2014). In MUDTv1, it is applied in a single instance only.

6.4.4.17 Other relations

6.4.4.17.1 Punctuation: punct

This relation is used for any and all punctuation dividing phrases, clauses and sentences. The following rules have been applied:

i. Sentence-end punctuation (.?!) attaches to the sentence root.
ii. Commas, colons, semi-colons and m-dashes (and equivalent characters) that separate clauses and phrases attach to the previous word, phrase or clause ...
iii. ... unless they appear in pairs, in which case they surround the word, phrase or clause in question.
iv. Paired punctuation marks (quotes, brackets, parentheses) are attached to the word, clause or phrase they surround...
v. ... unless this would lead to non-projective dependencies, e.g. when a quote is preceded by a full stop attached to the root. In such cases, such paired punctuation marks are attached to the sentence-end punctuation to the left.

Punctuation marks which stand for content words (typically nouns) like currency units or units of measure (€, £ or °) are treated as nouns.

6.4.4.17.2 Other dependency: dep

In MUDTv1, this relation (which might as well be named buggeredifIknow) is used for words that do not seem to be in a governor-dependent relationship with anything else. There are only two instances of dep being used in MUDTv1, one for a word which appears to have no relation to anything else in the sentence in either syntactic or semantic terms, and one for an apparent typo where a word is repeated twice for no obvious reason.

6.5 Data selection

6.5.1 Goals

The purpose of MUDTv1 is twofold: first, it is to provide the data for a quantitative analysis of constituent order variation in Maltese; secondly, it will be used to train and test automated parsers of Maltese. For both these, it is desirable that the selection of texts should reflect the entire spectrum of Maltese as much as possible. As for the former purpose, to put it bluntly, one should avoid the all-too-common scenario where a treebank composed of a single text type is used to make judgments about a language as whole.[10] For the second purpose, one would expect that the parser developed with

10 To give but one example: Heylen's quantitative study of German constituent order (Heylen 2005) is based on the NEGRA corpus which consists solely of newspaper texts from the *Frankfurter Rundschau*

MUDTv1 data will be used for various types of texts; consequently, it would be advantageous for the parser performance to expose it to the same kind of varied data it is likely to encounter.

With that in mind, the primary goal I had in the selection of texts for MUDTv1 was to cover as much of the spectrum of Maltese (as defined in Chapter 1, section 1.3.2.1) as possible, in as balanced way as possible. The secondary goal was, as with all corpus building, to obtain as much data as possible. This one had a very important practical consequence: as this is the first attempt at a treebank of Maltese, most of the work on syntactic annotation would have to be done manually. Consequently, the desire to end up with as much data as possible had to be balanced against the practicality of what could be achieved with a manageable amount of effort within a reasonable time frame. This required careful planning which in turn required setting a target number of sentences. For this, I reviewed the treebanks in UD v1 and their sentence counts (Nivre, Ginter et al. 2014) to find the smallest UD v1 treebank and aimed for a satisfying number higher than that, preferably with a more favorable sentence-to-speaker ratio. I finally settled on 2000 sentences as the goal for MUDTv1: this is higher than Buryat, Irish, Kazakh, Latvian, Tamil, Ukrainian and Uyghur, and on par with Hungarian which has at least 37 times as many native speakers.

6.5.2 Treebank composition

The primary goal of data selection for MUDTv1 was to be achieved in terms of ensuring the equal representation of the four text types from which *BCv3* was sourced, as the texts in MUDTv1 would come from *BCv3* anyway. Additionally, however, I decided to further divide each of the four categories into two subtypes based on available data based on internal or external criteria: newspaper into news items and op-eds, fiction into novels and short stories, non-fiction into humanities and sciences (including encyclopedic and instructional texts) and parliament into debates and questions (Q&A).

For each of the 8 subtypes, texts were selected so that each subtype would contain approximately 250 sentences. One important condition had to fulfilled as well: each text had to be as self-contained as possible, so as to ensure that the discourse flow and information structure division of sentences is maintained. This is not problematic with regard to newspaper texts or parliamentary transcripts, but walking the line between this requirement and the desired number of sentences necessitated the inclusion of a number of texts that were not included in *BCv3*, especially in the fiction and non-fiction

(coli.uni-sb.de/sfb378/negra-corpus/, last consulted on February 28th 2018.). Aware of the limitations, Heylen nevertheless insists that "[these] newspaper texts cover many topical domains and are written by multiple authors, often with different backgrounds, so that the patterns we find in this type of data might well be representative for modern German usage in general" (Heylen 2005: 245). Oy vey.

text types. It also presented some problems with regard to books; in that case, entire chapters or equivalent divisions are considered self-contained texts.

The original approach as described above worked reasonably well up until the treebank was about 80% complete when it became obvious that to annotate the remainder of the planned parliament files would cost me the last vestiges of my sanity. In the interest of preserving the same and completing this dissertation, I decided to remove parliament as a distinct text type and reshuffle the already annotated texts (plus adding as little as could be gotten away with) into a new text type. This text type I tentatively named quasi-spoken and it includes parliamentary debates and Q&A transcripts, as well newspaper interviews. The idea here is that while I did not set out to make any claims regarding spoken Maltese (see Chapter 1, section 1.3.2.1), it would nevertheless be of use to include texts that originated as spoken, if only for variety and rudimentary comparison.

Each of the files received a file code in the format AA_BBXCC where AA stands for a sequential number across all files, BB stands for a sequential number for a particular subtype, X stands for the text type (J for newspaper, P for parliament, F for fiction and N for non-fiction) and CC is a sequential number representing the subtype across all files. For example, the code 51_02N10 indicates that the file in question is text number 51 (51_), 2nd (02) in this particular subtype of the non-fiction text type (N) where the subtype – in this case, humanities – is the 10th (10) subtype in total. This is the original scheme which I did not modify during the reshuffling above and so the quasi-spoken text type does not receive its own alphabetical code.

Table 6.8 provides a summary description of the composition of MUDTv1, Tables 6.9 and 6.10 contain the full list of files with their codes, original file names and sentence counts arranged by text type and subtype. These require one note and one clarification: first, since syntactic annotation presupposes part-of-speech tagging, the primary source of data was the set of manually tagged files, one half of which was supplied by Albert Gatt (see Chapter 6, section 6.4.1.4). As a result, some of those files ended up in MUDTv1 and while their text type and subtype could easily be established, their actual origin could not be. In the tables below, the file names for such documents are given as MLRS with a number as provided to me as identification.

As the reader's keen eye will notice, the numbers in the file codes as given below are not sequential throughout. This is due a number of minor changes in the composition of the treebank made during the annotation process and the major reshuffling described above. In some text types, there were to be a few more files or even text subtypes. In others, the original set of files contained a few errors in sentence splitting; when those were removed, the final number of sentences was well below the targeted 250 and so a new file had to be added.

In its final form, MUDTv1 comprises 2074 sentences and 44,162 tokens total. And this brings me to my final point: one might ask if a treebank of this relatively modest size is sufficient for the analysis of any syntactic phenomenon, let alone constituent order and information structure. Several replies come to mind, most of them not fit

for print, but I will give one that is: there have been peer-reviewed works published which examined the same phenomenon in other languages and did it with much less. Taylor and Pintzuk (2012) examine the pragmatically determined position of objects in Old English on 394 clauses (actually object tokens, Taylor and Pintzuk 2012: 52), Tonhauser and Colijn analyze the constituent order in Guaraní on 2800 Guaraní words (Tonhauser and Colijn 2010: 259) and there exist a number of analyses of Old High German constituent order based on the same small corpus: for example, Cichosz 2010 lists 1505 clauses (Cichosz 2010: 52) as the count for her corpus of Old High German texts. So to answer the question, I say yes, providing reasonable caution in interpreting the results is exercised, 2074 sentences and 44,162 tokens is quite sufficient.

Text type	Subtype	Sentence count
newspaper	news	239
	op-eds	240
	Subtotal	479
quasi-spoken	newspaper interviews	280
	parliament: debates and Q&A	294
	Subtotal	574
fiction	short stories	246
	novel chapters	251
	Subtotal	497
non-fiction	humanities	249
	science, encyclopedic and instructional	275
	Subtotal	524
	Total	2074

Tab. 6.8: MUDTv1 composition: summary

Text type	Subtype	File code	File	Sentence count
newspaper	news	01_01J01	MaltaRightNow 58-99838391 (2012-06-26)	19
		02_02J01	INewsMalta (2015-02-26)	43
		03_03J01	MaltaRightNow 19-99837404 (2012-05-20)	13
		04_04J01	MLRS 01	11
		05_05J01	L-Orizzont 100366	25
		06_06J01	MLRS 17	9
		07_07J01	MLRS 27	20
		08_08J01	MLRS 28	13
		09_09J01	MLRS 29	19
		10_10J01	Il-Ġens Illum 7938	23
		11_11J01	MLRS 33	13
		12_12J01	MaltaRightNow 20-99825435 (2011-01-11)	15
		13_13J01	MaltaRightNow 19-99827053 (2011-03-25)	16
			Subtotal	239
	op-eds	14_01J02	L-Orizzont 95698	37
		15_02J02	MLRS 09	21
		16_03J02	MLRS 16	18
		17_04J02	MLRS 07	73
		18_05J02	MLRS 22	37
		19_06J02	MLRS 35	20
		20_07J02	It-Torċa 7677	34
			Subtotal	240
quasi-spoken	newspaper: interviews	21_01J03	MLRS 15	51
		22_02J03	Illum, interview (2008-03-30)	149
		23_03J03	Newsbook, sports (2014-08-15)	17
		23b_04J03	Illum, interview (2006-12-17)	63
			Subtotal	280
	parliament: debates	30_01P05	File 20150218_060d_kon.docx	219
	parliament: Q&A	38_02P06	File 20020626_762d_par-QA	75
			Subtotal	294

Tab. 6.9: MUDTv1 composition: text types newspaper and parliament

Text type	Subtype	File code	File	Sentence count
fiction	short stories	39_01F08	Għidli Mitejn, JAKE JEW NATHAN?	14
		40_02F08	Għidli Mitejn, MA KINITX KELMA	20
		41_03F08	Għidli Mitejn, NORMALI	22
		42_04F08	Għidli Mitejn, IN-NATURA U JIEN	26
		43_05F08	Għidli Mitejn, IL-KUĠINA	26
		44_06F08	Għidli Mitejn, PERPETUUM MOBILE	24
		45_07F08	Għidli Mitejn, ALTAF NAIFEH	18
		46_02F08	Clare Azzopardi - Danubju	96
			Subtotal	246
	novel chapters	47_01F09	2012 John A. Bonello - It-Tielet Qamar (unnumbered first chapter, p. 1-4)	76
		48_02F09	1998 Rena Balzan - Ilkoll ta' Nisel Wieħed (first section of chapter 4, p. 41-44)	66
		49_03F09	2012 Trevor Żahra - Il-Ġenn li Jżommni f'Sikkti (first two sections of chapter 12, p. 300-304)	73
		49b_04F09	2011 Loranne Vella - Magna Mater (chapter 31, p. 243-244) (not in BCv3)	36
			Subtotal	251
non-fiction	humanities	50_01N10	MLRS 31	128
		51_02N10	1999 Lawrence E. Attard - L-Emigrazzjoni maltija (not in BCv3)	34
		52_03N10	2011 Charles Briffa - Il-Varjetajiet tal-Malti (Part I, chapter 2, section "Id-djalett", p. 18-19) (not in BCv3)	33
		53_04N10	2012 Guido Lanfranco - Drawwiet u tradizzjonijiet maltin (chapter 5, introduction and section "Qabel it-tqala", p. 59-61) (not in BCv3)	54
			Subtotal	249
	science, encyclopedic, instructional	54_01N11	Il-Ġens Illum 1000	14
		55_02N11	Wikipedia, entry "Alġebra astratta"	72
		56_03N11	2001 - Peter A. Caruana - Għarfien il-pjanti (entry "Capsicum", p. 92-97)	87
		57_04N11	Manual "L-Użu tal-Malti fil-Kompjuter" (section "L-Unicode (UTF8)", p. 39-41) (not in BCv3)	17
		58_05N11	Wikipedia, entry "Materja skura"	85
			Subtotal	275

Tab. 6.10: MUDTv1 composition: text types fiction and non-fiction

6.5.3 Manual annotation

As MUDTv1 is the first attempt at creating a Maltese treebank, the bulk of the work on the annotation had to be performed manually. This is, naturally, a task that could theoretically be accomplished using any decent text editor, but efficiency, consistency and self-preservation suggest that the use of a dedicated software tool would be advisable.

There exist a number of tools that are or can be used for syntactic annotation, either ones designed specifically for this purpose, such as TrEd[11] and Arborator,[12] or multi-purpose annotation environments like brat (Stenetorp et al. 2012) and WebAnno (Yimam et al. 2014). Having surveyed the available options, I found that nearly all of these tools had serious issues: for one, they are powerful, but complex and thus difficult to use (TrEd is a very good example), but more importantly, they are anything but user friendly (such as Arborator with its clumsy mouse-controlled interface). The task at hand essentially boils down to manipulating and connecting objects in two dimensions, surely there are better options of implementing this functionality. One such option involves the use touchscreen technology on mobile devices and this is a solution I focused on from the outset. I enlisted the help of an expert and together we devised and developed a tool that addresses the shortcoming of existing tools for syntactic annotation.

As the platform, we chose iOS for its intuitiveness, reliability and general design philosophy: iOS apps are notorious for doing exactly what it says on the box and doing it perfectly. This was the guiding principle behind the effort which resulted in the iOS app called PosTagger. In specifics, PosTagger was designed as an iPad-first app with four basic requirements:

I. Use plain-text UTF-8 encoded files as input, parameters and output.
II. Enable manual customized part-of-speech tagging.
III. Enable manual customized chunking.
IV. Enable manual UD v1 syntactic annotation.

The idea behind this is to take the vertical text files used to compile *BCv3* without any modification, import them into the app, process them and obtain human-readable and easily-processible text files as output. The customization is also implemented by means of plain-text files: for example, for part-of-speech tagging, the list of tags is supplied to the application as a simple text file with one tag per line.

The actual annotation would then be performed by means of tactile interaction: for part-of-speech tagging, this would involve tapping on a token and then selecting a tag from the list of tags; once the selection is made, the focus automatically moves to the next token. The app of course remembers the choice made for each token and when

11 ufal.mff.cuni.cz/tred/ (last consulted on February 28th 2018)
12 arborator.ilpga.fr/ (last consulted on February 28th 2018)

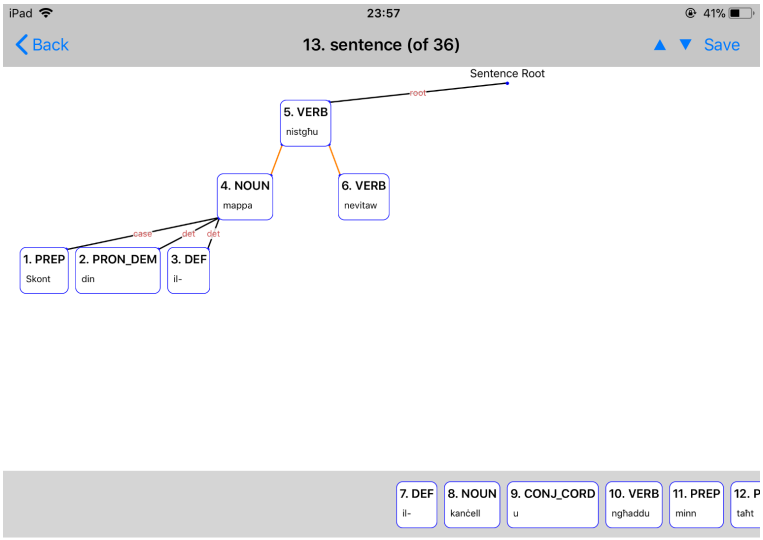


Fig. 6.1: PosTagger: Syntactic annotation

that token is encountered again, the previously selected tag is filled out automatically and the user can either accept it or make a correction.

For syntactic annotation, the process is a little more complicated: firstly, syntactic annotation presupposes part-of-speech tagged and chunked text. The chunking is an artifact of the original design where the intention was to manipulate chunks, not tokens. This was quickly abandoned, but the requirement remains, except that now chunks are by default indentical to tokens. Once a part-of-speech-tagged file is read into PosTagger, the user can select any token in the main interface and the sentence to which the token belongs appears in the syntactic annotation interface. The sentence is displayed as a horizontal list of chunks/tokens with their part-of-speech tags at the bottom of the screen, with tokens and tags contained in small white rectangles with blue perimeter. The user can then manipulate the rectangles by tapping and moving them; once a rectangle is moved sufficiently close to another rectangle, the stationary rectangle changes color to green, indicating the rectangles can be connected. When the moved rectangle is released, it locks up in place and a line appears connecting the two rectangles. The rectangle positioned higher is the governor of the rectangle below; consequently, a rectangle can only connect to a rectangle that is above them or to the top of the window where it connects to a point labeled "Sentence Root" (see Figure 6.1). When rectangles are first connected, the lines that connect them are orange in color and unlabeled. To label them, the user taps the rectangle and in the pop-up that appears (Figure 6.2), selects the applicable relation. Once a relation is selected, the line turns black and the relation appears on the line. The relation can be changed at any time following the same procedure.

Fig. 6.2: PosTagger: Relation selection

One rectangle must always be connected to ”Sentence Root”, but this is the only such limitation; in principle, any rectangle can connect to any other (but, naturally, only one), even in a manner that creates non-projective dependencies. There can even be multiple rectangles connected to ”Sentence Root”, although in this case, the top of the screen is colored red to warn that this is an error. Once connected, rectangles can be moved at a whim; if they already have a label assigned, it is retained. This also applies to entire catenae: if a user would tap, hold and move the rectangle containing the token *mappa* in Figure 6.1, the token and its dependents would move as a single unit.

Once the user is satisfied with the annotation, they can use the ”Save” button to save the document or use the arrows in top right corner to move to the next (or return to the preceding) sentence (in which case the document is saved automatically) until the entire file has been annotated. The file can then be exported using the default iOS functionality to any cloud storage or sent via email, iMessage or any other app.

While PosTagger accepts vertical text as input and uses the same format for the output of part-of-speech tagging and chunking, for syntactic annotation, a different format had to be employed. This is necessitated by the internal workings of the app which requires a data storage format capable of capturing the two-dimensional relations involved. To satisfy the requirement for output as outlined above (UTF-8 encoding, plain text, human readability), XML was chosen as that format which for the purposes of export receives the file extension *.ptg (for ”parsed and tagged”). The following is an example of a PTG file:

```
<doc orig="dummyfilename.vrt.ptg">
<sentence pos="0">
<group dep="root">
<chunk pos="1" type="PRON_INT" relation="root">
<token tag="PRON_INT">X'</token>
</chunk>
<group dep="cop">
<chunk pos="2" type="PRON_INT" relation="cop">
<token tag="PRON_INT">inhu</token>
</chunk>
</group>
<group dep="nsubj">
<group dep="det">
<chunk pos="3" type="DEF" relation="det">
<token tag="DEF">l-</token>
</chunk>
</group>
<chunk pos="4" type="NOUN_PROP" relation="nsubj">
<token tag="NOUN_PROP">Unicode</token>
</chunk>
</group>
<group dep="punct">
<chunk pos="5" type="X_PUN" relation="punct">
<token tag="X_PUN">?</token>
</chunk>
</group>
</group>
</sentence>
...
</doc>
```

The root element of a PTG file is <doc>; each <sentence> element is then a direct child of the root. The element <group> represents the rectangles, i.e. the tokens, which are contained within the <token> element encapsulated in the <chunk> element. If a token has no other dependent, the <chunk> element is the direct dependent of <group>; any dependents of any token are encapsulated in a child <group> element. The dependency relations thus hold between groups, but are marked as both the dep attribute of <group>, as well as the relation attribute of <chunk>. The order of sentences is recorded as the pos attribute of the <sentence element>, the order of tokens is recorded using an attribute of the same name on the <chunk> element; this ensures that the correct order of tokens is preserved in case of non-projective dependencies. This is an artifact of the original design where chunks, not tokens, were

to be the units of syntactic annotation and the existence of the type attribute in the <chunk> element is further evidence for it: originally, it was to contain the chunk type (e.g. NP for "noun phrase"), but now it contains the part-of-speech tag and thus the information that the tag attribute in the <token> element also provides.

The PTG format is simple and provides all the flexibility of XML. The exported PTG files can thus be queried with XPath or converted with XSL; the latter is how PTG is converted to CoNLL-U for further enrichment and to be used in corpus management software.

6.5.4 Corpus management and querying

For visualization and analysis purposes, ANNIS3 (Krause and Zeldes 2016) was selected for its flexibility, adaptability and ease of use. The CoNLL-U files were converted to the native ANNIS format using the Pepper platform (Zipser and Romary 2010) with only the XPOSTAG and the dependency layers retained so as to work around certain limitations when it comes to querying and visualizing dependencies. The ANNIS instance is available publicly at bulbul.sk/annis-gui-3.4.4.

In addition to this, the full set of files was converted to the brat annotation format (Stenetorp et al. 2012), which is also the standard format for UD visualization. The individual HTML files are available at bulbul.sk/bonito2/treebank (login name: guest, password: Ghilm3).

7 Dominant constituent order and its variations in Maltese: A quantitative analysis

7.1 Introduction

With the metalanguage (Chapter 1), data (Chapter 5 and 6) and methodology (Chapter 4) established, I will now proceed to answering the research questions in the order in which they were asked, starting with the question of what is the statistically dominant constituent order in Maltese, both in general terms, as well as in specific types of clauses. Subsequently, I will analyze the variation in constituent order (i.e. the fact that some types of clauses exhibit dominant order opposite to that of the majority of clause types and Maltese as a whole or they exhibit no dominant order at all) and attempt to account for it. And finally, I will use the findings to provide a typological characterization of the constituent order in Maltese.

The analysis here was conducted using data obtained from the ANNIS corpus management software (Chapter 6, section 6.5.4) or exported from PTG files (Chapter 6, section 6.5.3) and processed using the R software environment. The full list of ANNIS queries, the data extracted from PTG files and the R scripts used to process them are included in Appendix C.

7.2 Basic statistics

7.2.1 Sentence length and complexity

I begin with a brief aside to provide basic information on the sentences in MUDTv1 and to check three general observations (or rather impressions). First, there is the question of sentence length in general: in the process of manual annotation, the differences between text types in the number of tokens per sentence were often felt to be quite pronounced; this is doubly true of newspaper texts annotation during which many an expletive was hurled in the general direction of Maltese journalists for the long and cumbersome sentences they produce. To confirm whether such differences are actually real, I extracted the pos ("position") ID number of the last token in every sentence which captures the total number of tokens in a sentence,[1] calculated the mean, the standard deviation (SD) and the standard error of the mean (SE) for each text file (Table

1 In POSTagger PTG files, pos is zero-indexed, so the actual number of tokens is last token pos + 1. This has been corrected for the graphs below.

Text type	Total sentence count	Mean sentence length (tokens)	SD	SE
newspaper	479	26.8	13.25	0.60
quasi-spoken	574	21.6	16.18	0.67
fiction	497	16.9	11.07	0.49
non-fiction	524	24	13.19	0.57
MUDTv1	2074	22.3	14.09	0.31

Tab. 7.1: MUDTv1: Mean sentence length by text type

7.1)[2] and then plotted the numbers (along with 95% confidence intervals) in Figure 7.1.

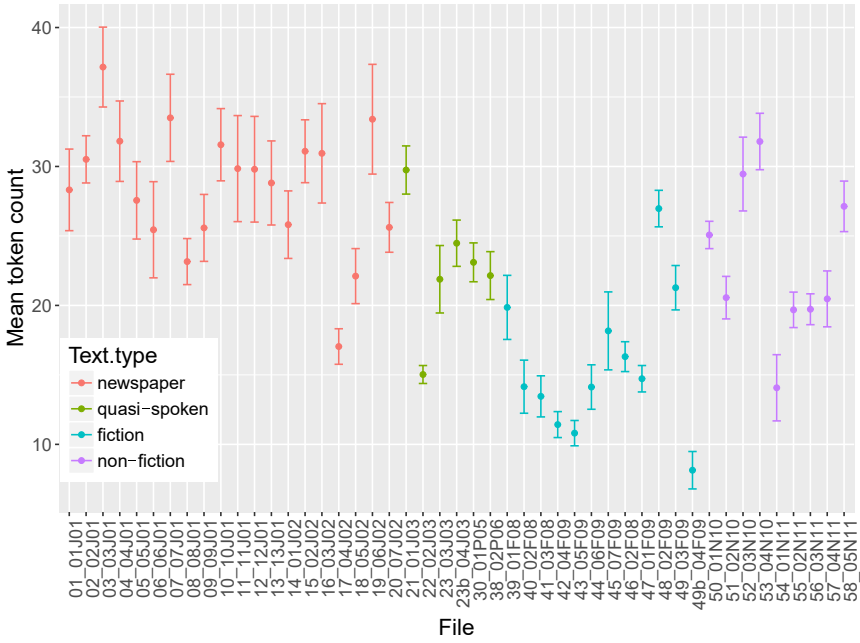


Fig. 7.1: MUDTv1: Mean sentence length by file and text type

The numbers in 7.1 show clear difference between the text type fiction and the three other text types and thus confirm the observation; the breakdown by MUDTv1 text type and file only underscores this. These figures also confirm previous findings

2 For this calculation, I used the method described in bit.ly/1nn8uw0 (last consulted on February 28th 2018).

by Fenech (1978: 222) regarding sentence length in journalistic texts as compared to other text types:

> "The news report, the editorial, and the article have 28, 27, and 21 words per sentence respectively. These are all higher than the average number of words per sentence in literary Maltese (18 words) and in the spoken language (5 words)..."

In MUDTv1, the newspaper text type includes both news reports (subtype code 01) and editorials (subtype code 02, see Table 6.9). The average sentence length for the former stands at 29.6 and at 24 for the latter, again closely mirroring Fenech's findings, despite a near 40-year difference between the samples.

The second general observation involves the existence of a substantial difference in sentence length between those texts that originated in writing tout court and those that have been transformed (in one way or another) from spoken language. As noted in Chapter 1, section 1.3.2.1, this work examines Maltese as represented by written texts in *BCv3* and MUDTv1 and makes no claims about the spoken language. Nevertheless (as discussed in Chapter 6, section 6.5.2), there are two subcategories of texts that originated as transcriptions (albeit heavily edited ones) of speech: newspaper interviews and transcripts of parliamentary proceedings, subsumed under the quasi-spoken text type. These can be contrasted with those texts that originated purely in writing (newspaper, fiction and non-fiction text types) and it would make sense to see if there is a contrast comparable to that seen above. Table 7.2 and Figure 7.2 serve to check and confirm this.

Text type	Mean sentence length (tokens)	SD	SE
written	22.6	13.20	0.34
quasi-spoken	21.6	16.18	0.67

Tab. 7.2: MUDTv1: Mean sentence length in written and quasi-spoken texts

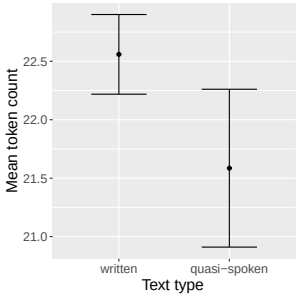


Fig. 7.2: MUDTv1: Mean sentence length in written and quasi-spoken texts

And finally, the third observation relates to the complexity of sentences in MUDTv1, expressed in terms of the number of dependent clauses (as defined in UD v1, see also section 7.2.2.1 below). These have been extracted from the PTG files and ploted in Figure 7.3 as absolute numbers for each of the 2074 sentences in MUDTv1 arranged by text type. Additionally, the position of each dependent clause is recorded as either pre-root (negative numbers) and post-root (positive numbers), resulting in a pyramid-like plot.

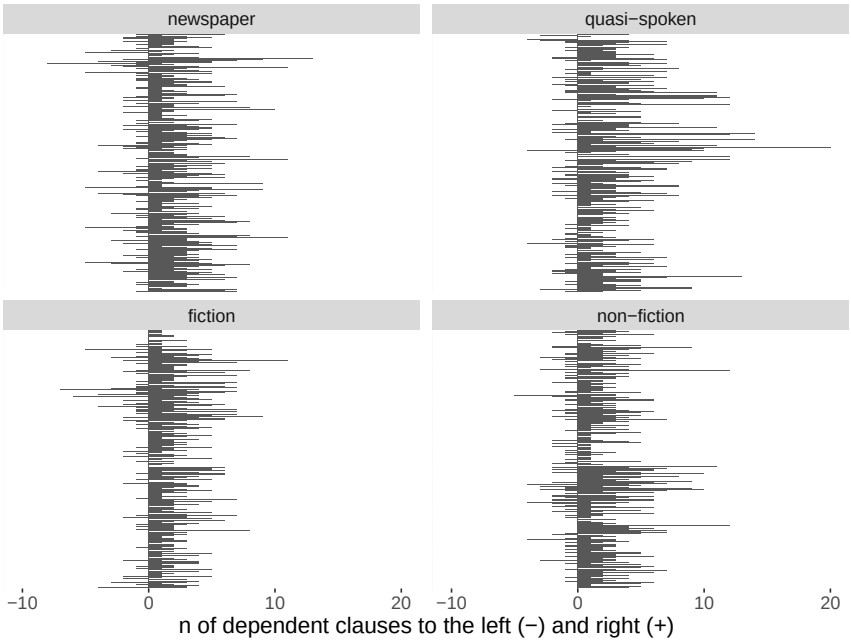


Fig. 7.3: MUDTv1: Sentence complexity by text type

This plot shows differences in sentence complexity between text types, once again positioning fiction as the outlier (see also Figure 7.4), but it also provides a classification of Maltese in terms of ordering of clausal dependents in complex sentences: Maltese, as apparent from the MUDTv1, is consistently right-branching.[3]

3 This is, nota bene, only a subset of the classification established by Dryer (1992) and should not be taken to be indentical with it.

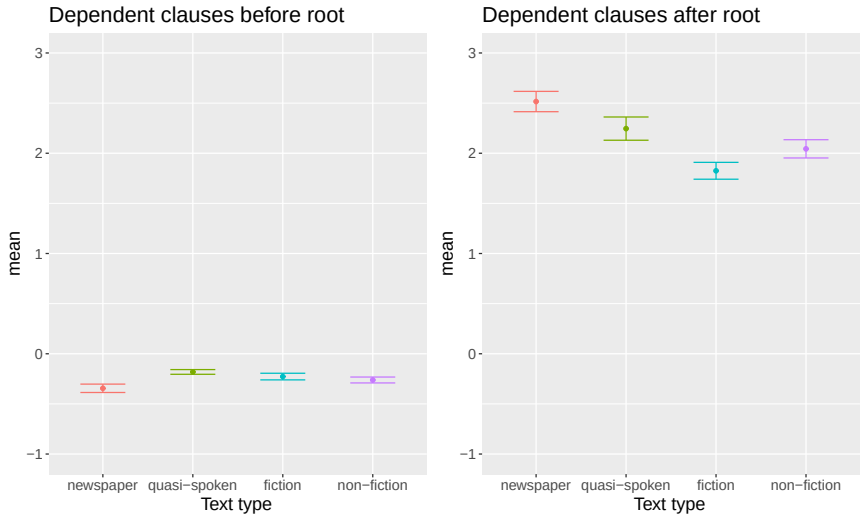


Fig. 7.4: MUDTv1: Sentence complexity by text type

All the findings above provide further justification for including as many and as varied types of texts in MUDTv1 as possible: relying only on one text type – say, journalistic texts – as many treebanks do (cf. Heylen 2005, but also Nivre, Agić et al. 2017) would provide a very skewed picture of Maltese. And while representativeness may be a pipe dream for Maltese corpus linguistics, balance should be a priority.

7.2.2 Clause types

7.2.2.1 General

The following analysis will consider constituent order from the point of view of two classifications of clauses employed here so far:

I. UD clause types (Chapter 6, Table 6.5)
II. Clause types by root (Chapter 6, section 6.4.4.1)

In this section, basic statistics for both classifications are provided.

7.2.2.2 UD clause types

Table 7.3 below contains an overview of UD clause types in MUDTv1 (as per UD v1) and the count of their occurrences.

Clause type	Number
Main clauses	2074
acl	1318
advcl	843
xcomp	1375
ccomp	684
parataxis	185
conj	871
csubj	23
Total clauses	7373

Tab. 7.3: MUDTv1: UD clause types

This list includes all main clauses (root) and their direct or indirect dependents with the labels listed above, save for conj; in their case, only those catenae with a head marked as conj were taken into account that were tagged VERB, VERB_PSEU, PART_ACT, PART_PASS and HEMM or had a cop or nsubj as a dependent. Contrary to Table 6.5, the list relation is not included here, as in MUDTv1, this relation is not used for clauses.

7.2.2.3 Clause types by root

The second classification of clauses used here is that by their structure, i.e. their root (Chapter 6, section 6.4.4.1). Since it is the order of the predicate (root) and its core nominal dependents (nsubj, nsubjpass, dobj and nmod:obj) that is the primary focus of this analysis, only those clauses that contain at least one of the latter group will be considered for the purposes of the analysis attempted here. Table 7.4 contains an overview of all such clauses and their counts by UD clause type.

UD clause type	Verbal nsubj	Verbal nsubjpass	Verbal dobj	Verbal nmod:obj
main	699	73	408	74
acl	146	35	246	61
advcl	201	31	214	21
xcomp	40	6	371	75
ccomp	278	44	136	28
parataxis	67	1	23	6
conj	143	25	215	36
csubj	8	0	4	0
Total	1582	215	1617	301

UD clause type	Copular nsubj	Existential nsubj
main	231	54
acl	23	13
advcl	47	25
xcomp	7	10
ccomp	99	44
parataxis	10	3
conj	45	20
csubj	1	0
Total	463	169

Total nsubj	2205
Total nsubjpass	215
Total dobj/nmod:obj	1918

Tab. 7.4: MUDTv1: Clauses containing core dependents by root (columns) and UD clause type (rows)

7.3 Constituent order in MUDTv1 by the numbers

7.3.1 Overview

In this section, I offer a quantitative analysis of constituent order configurations in MUDTv1 using tables and graphs. This analysis will employ the standard convention of using the abbreviations S, V and O with one major distinction: here, V will stand for any type of predicate, not just a verbal one. Further deviations from this convention will be noted as necessary.

As outlined in the research questions, the constituent order in MUDTv1 will be examined from the point of view of the SV/VS and VO/OV dichotomies (Dryer 1997 and 2013a). Before doing that, however, I provide below an overview of constituent order according to the Greenbergian six-way classification, if only for comparison's sake. There are only 472 verbal clauses with both a subject and an object; Figure 7.5 plots the distribution of all six types regardless of UD clause type or clause root, while Table 7.5 provides the same information in absolute as well as relative numbers.

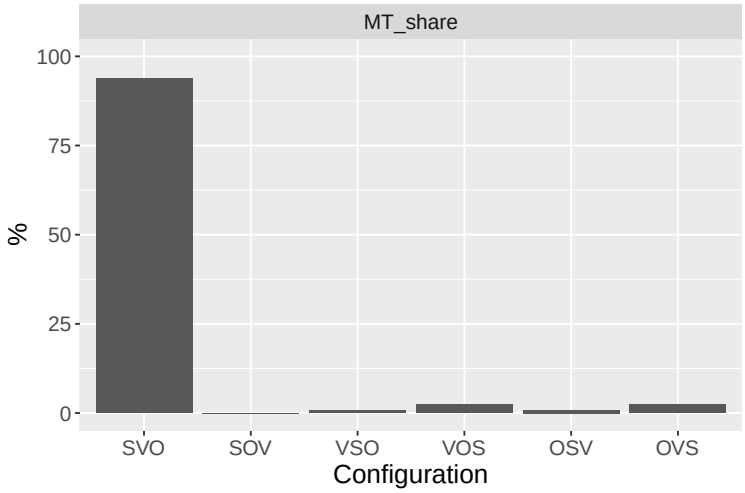


Fig. 7.5: MUDTv1: Constituent order – Greenbergian classification

Configuration	Count	%
SVO	443	93.86%
SOV	0	0.00%
VSO	3	0.63%
VOS	11	2.33%
OSV	4	0.85%
OVS	11	2.33%
Total	472	100%

Tab. 7.5: MUDTv1: Constituent order – Greenbergian classification

This data provides a clear picture of Greenbergian classification of Maltese, as well as further justification for Dryer's SV/VS and VO/OV typology: Greenbergian typology only has 472 data points to work with; using Dryerian typology, the data sample expands more than five-fold for subjects (2420 total) and four-fold for objects (1918).

In what follows, I will go on to provide a more fine-grained classification, beginning with a general overview of the distribution of constituent order configurations in MUDTv1 across clause types classified by root (Figure 7.6).

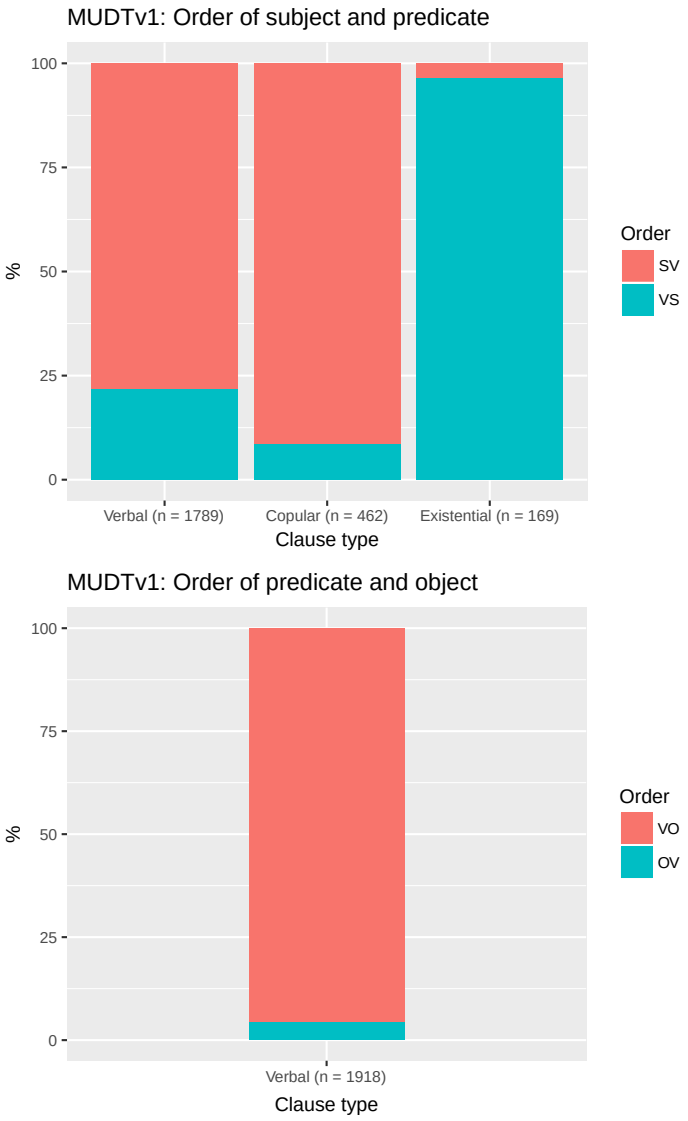


Fig. 7.6: MUDTv1: Constituent order by clause type – overview

This is a finer, yet still somewhat coarse analysis, where S combines the counts for both nsubj in verbal active clauses, copular clauses and existential clauses and nsubj-pass in verbal passive clauses. Likewise, O combines the counts for relations dobj and nmod:obj for all verbal clauses.

A similar analysis could be attempted for non-copular verbless clauses by examining the position of the expletive subject in relation to its head. There are 21 such constructions in MUDTv1 and in all of those, expl (whether a PRON_PERS or KIEN) invariably precedes its head. Non-copular verbless clauses will therefore not be included in the analysis here; neither will, for obvious reasons, non-expletive subjectless clauses.

The following sections provide the advertised fine-grained analysis first by root and then by UD clause type, with one exception: due to their low number, clauses of the csubj type will be excluded from this analysis.

7.3.2 Verbal clauses

7.3.2.1 Introductory remarks

The analysis below is primarily based on UD relations; this is true of both the classification of clauses, as well as the treatment of constituents. Consequently, nsubj and nsubjpass are treated separately; by extension, so is nmod:agent. Objects (dobj, nmod:obj and iobj), on the other hand, are treated equally, regardless of whether the clause is active or passive; their analysis thus relies on the verb's valency frame.

7.3.2.2 Order of active subject and predicate

7.3.2.2.1 General

Figure 7.7 plots the distribution of SV and VS configurations in active verbal clauses across UD clause types, Table 7.6 provides the same information while also adding absolute numbers. As above, this graph and this table are only based on data for active clauses, i.e. the S here stands for nsubj only.

The 8 counts of verbal csubj have been excluded from this overview. For completeness' sake, I add here that the distribution of SV and VS configurations in these clauses is 4-4 and thus no dominant constituent order can be established.

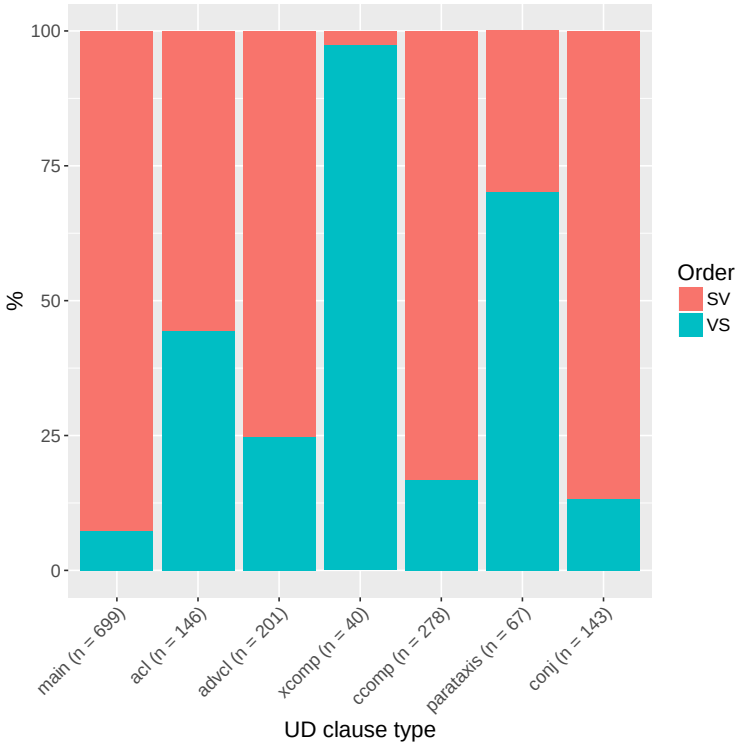


Fig. 7.7: MUDTv1: Order of subject and predicate in active clauses by UD clause type

UD clause type	Order	Count	%
main	SV	647	92.56%
	VS	52	7.44%
acl	SV	81	55.48%
	VS	65	44.52%
advcl	SV	151	75.12%
	VS	50	24.88%
xcomp	SV	1	2.50%
	VS	39	97.50%
ccomp	SV	231	83.09%
	VS	47	16.91%
parataxis	SV	20	29.85%
	VS	47	70.15%
conj	SV	124	86.71%
	VS	19	13.29%
Total	SV	1255	79.73%
	VS	319	20.27%

Tab. 7.6: MUDTv1: Order of subject and predicate in active clauses by UD clause type

Leaving aside the question of acl where no dominant order could be established, the data above might lead us to conclude that the dominant order of predicate and subject in verbal clauses is SV except for xcomp and parataxis where VS is the dominant configuration. Before we do that, however, let us take a closer look at both these clause types.

7.3.2.2.2 VS in active xcomp

As evident from the summary Table 7.7 below, active verbal xcomp seem to exhibit VS as the clearly dominant order.

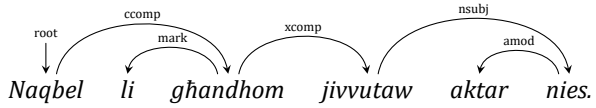
UD clause type	Order	Count	%
Active xcomp	SV	1	2.50%
	VS	39	97.50%

Tab. 7.7: MUDTv1: Dominant VS in active xcomp

A cursory look at the list of the 39 VS xcomp clauses reveals a pattern and a problem with this analysis: each and every one of these clauses is an instance of a verbal chain. As noted earlier (Chapter 6, section 6.4.4.4.3, section 6.4.4.8.1, see also Stolz 2009 and Fabri and Borg 2017), verbal chains are sequences of two or more verbs (including KIEN) or pseudoverbs which share a single subject and where the final verb in the chain is the lexical one and the only one capable of bearing nominal core arguments other

than the subject. In MUDTv1, each verb is analyzed as a separate xcomp clause (1) with items breaking the chain (Stolz 2009) assigned as dependents to the respective xcomp.

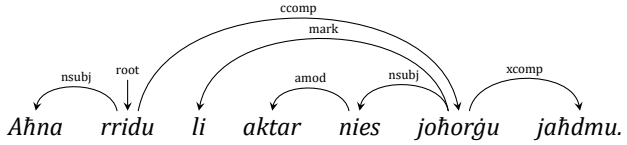
(1) *Naqbel li għandhom jivvutaw aktar nies.*
I agree COMP they have they vote more people
'I agree that more people should vote.'



[MUDTv1: 22_02J03]

This is the chief problem with the analysis of xcomp as implemented in MUDTv1: despite the insertion of some elements, verbal chains act as a single syntactic unit (Fabri and Borg 2017: 80-82), especially with respect to subjects. In other words, subjects can appear either before the first verb in the chain (2) or after the last one (1), but not between the links in the chain.

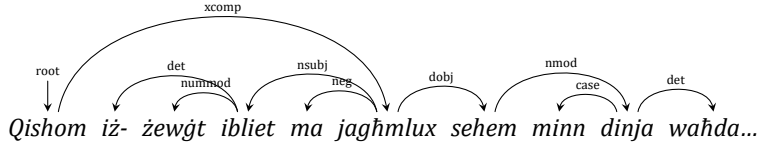
(2) *Aħna rridu li aktar nies joħorġu jaħdmu.*
we we want COMP more people they go out they work.
'We want more people to go out and work.'



[BCv3: 20131120_092d_par]

Consequently, the searches used in obtaining the numbers in Table 7.6 and Figure 7.7 actually represent two different and not at all complementary types of information: for SV, looking for a nsubj preceding an xcomp will match only those verbal xcomp that are not a part of a verbal chain; in MUDTv1, the solitary example is (3) with the VERB_PSEU *qis-* "to be like" as the head of the clause.

(3) *Qishom iż- żewġt ibliet ma jagħmlux sehem minn dinja waħda...*
like-3PL DEF two city.PL NEG they make-NEG part from world one-F...
'It's like the two cities are not a part of one world.'



[MUDTv1: 48_02F09]

For VS, on the other hand, looking for a nsubj following a xcomp only collects verbs that are the last in a verbal chain and have a nsubj after them. The numbers for VS xcomp clauses above therefore represent entire verbal chains and their actual syntactic role (UD relation) is that of the first verb in the chain.

That these are the only two types of xcomp clauses that feature a nominal subject is hardly surprising: by their definition, xcomp clauses inherit their subject from a higher clause. Those xcomp clauses that actually have a nsubj dependent are therefore special cases: in MUDTv1, these are dependents of *qis-* (3) and verbal chains. The former is a clear-cut case, although perhaps necessitating the revision of the status of *qis-* as a VERB_PSEU and possibly its syntactic role, a task which is outside of the scope of this work. As for verbal chains, the fact that its constituent verbs are analyzed as separate xcomp clauses obscurs the fact that they opperate as a single unit and thus the only instance where an xcomp clause can take a nominal subject is when the subject follows the verbal chain. Consequently, the variation discussed here is actually no variation in dominant constituent order in xcomp clauses at all. Rather, it is deviation from the dominant constituent order in the UD clause types by which the verbal chains (of which these xcomp clauses are members) are governed. And so for example in (1), the noun phrase *aktar nies* is the subject of the ccomp headed by *għandhom*, just as in (2), the same noun phrase is the subject of the ccomp headed by *joħorġu*.

The 39 cases of VS xcomp have therefore been reinterpreted in this manner and the respective xcomp needs to be redistributed among the heads of the verbal chains they are members of. Those are listed in Table 7.8 and these numbers should be added to the respective entries (the VS row) in 7.6 and subtracted from the xcomp entry.

Head	Count	%
main	9	23.08%
acl	8	20.51%
advcl	6	15.38%
ccomp	8	20.51%
parataxis	5	12.82%
conj	3	7.69%
Total	39	100.00%

Tab. 7.8: MUDTv1: Head of active xcomp

If one were to do so, one would see the picture painted by Table 7.6 and Figure 7.7 largely unchanged, except for xcomp, which now, like the remaining verbal UD clause types save parataxis, counts among those with dominant SV order, although it contains only a single clause.

7.3.2.2.3 VS in active parataxis

Paratactic clauses in MUDTv1 also seem to exhibit VS as the dominant order of subject and predicate (Table 7.9).

UD clause type	Order	Count	%
Active parataxis	SV	20	29.85%
	VS	47	70.15%

Tab. 7.9: MUDTv1: Dominant VS in active parataxis

As outlined in Chapter 6, section 6.4.4.15.2, the parataxis relation is used in MUDTv in two scenarios:

I. Dependent clauses that are not in coordination or subordination to their head clause (henceforth: parenthetical clauses).
II. Speech verbs and speaker identification following reported speech (whether quoted or rephrased; henceforth: reported speech clauses).

This division is also reflected in the distribution of SV and VS configurations and the dominance of VS is largely due to the frequency of the second type of parataxis clauses in MUDTv1: as evident from Table 7.10, of the 47 parataxis VS clauses, 38 are reported speech clauses.

Construction type	Count	%
Reported speech (cited)	37	78.72%
Reported speech (rephrased)	1	2.13%
ngħidu aħna	5	10.64%
Other	4	8.51%
Total	47	100%

Tab. 7.10: MUDTv1: Dominant VS in active parataxis: Types

Further 5 clauses are actually instances of *ngħidu aħna*, literally "say we", whether on its own or modifying an appositional modifier. This is an fixed expression used in explanatory sentences meaning "that is to say" and as such, its structure is fossilized. The actual distribution of constituent order variation in parataxis clauses is therefore as laid out in Table 7.11:

Construction type	Order	Count	%
Parenthetical clauses	SV	20	29.85%
Reported speech (cited and rephrased)	VS	38	56.72%
ngħidu aħna	VS	5	5.97%
Other	VS	4	8.51%
Total		67	100%

Tab. 7.11: MUDTv1: Classification of constituent order variation in active parataxis

Discounting the straightforward case of the fixed expression, what we have here are two major types of parataxis clauses (parenthetical and reported speech), each with their own obligatory structure and each failing to reach the 66% threshold. This, incidentally, remains true even after one adds the five clauses from the xcomp category: 3 of those are reported speech clauses and 2 are parenthetical, bringing the share of those two types to 56.94% and 30.56%, respectively.

What then are we to make of the four examples in the "other" category? A quick analysis of their structure reveals that they also come in two types. The first type is exemplified by (4):

(4) *Issa nħallihom jaqbduni, x' aktarx jagħmluli*
now I let-ACC.3PL they capture-ACC.1SG, some rather they make-DAT.1SG
xi diskursata l- Kunsill tad- disgħa...
some lecture DEF council GEN-DEF nine...
'Now I will let them capture me, they might lecture at me, the Council of the Nine...'

[MUDTv1: 14_01J02]

Of the four clauses in the "other" category, three are of this type. The one that is not is a parenthetical enclosed in parentheses (5):

(5) *(wara kollox hekk tfisser il- kelma Maltija għerq)*
(after all thus she means DEF word Maltese-F għerq)
'(this is, after all, what the Maltese word *għerq* means)'

[MUDTv1: 53_04N10]

It becomes immediately obvious that we are looking here at pragmatically motivated deviation from the default SV: in (4), the nominal subject is superfluous and provided for clarification; this is doubly true of the three remaining clauses where the nominal subject is a pronoun whose referent is already denoted morphologically. For (4), this analysis is supported by *BCv3* where one obtains 62 hits for the search query /.fisser hekk/, 157 /hekk .fisser/. The conclusion to be drawn here is that these clauses should

fall into "parenthetical clauses" type and their VS order to be interpreted as pragmatically motivated deviation from the dominant SV order for parenthetical clauses: such clauses are, after all, in no actual dependent relationship to their governor and should thus be considered syntactically equivalent to main clauses.

This analysis possibly necessitates the revision of the conclusion above: it cannot be ruled out that the dominant VS order in parataxis clauses is merely an accident of the composition of MUDTv1, in that it is due to the fact that the *ngħidu aħna* idiom and reported speech clauses (both VS) predominate and together make up 62.7% of all parataxis clauses. The remaining four clauses, which push the VS count over Dryer's threshold, are parenthetical clauses and as such, they should perhaps not count as parataxis proper. This would then mean that no dominant order can be established for active parataxis clauses in general. Only an expanded treebank, one that takes this division into account, can definitively answer this question. For MUDTv1, however, the conclusion above concerning VS as the dominant constituent order in parataxis clauses stands.

7.3.2.2.4 No dominant order in active acl

The only clause type in MUDTv1 where dominant order of S and V cannot be established is acl. Table 7.12 summarizes the absolute and relative counts for the order of S and V in active clauses.

UD clause type	Order	Count	%
Active acl	SV	81	55.48%
	VS	65	44.52%

Tab. 7.12: MUDTv1: Order of S and V in active acl

With this number of clauses and their often complex structure, a manual analysis proved to be impractical, so instead, I decided to employ statistical modeling to determine what syntactic factors (in terms outlined in Chapter 1) and other factors (see Research Question 4) influence the ordering of subject and predicate in these clauses. First, I established four categories of possible relevant syntactic factors:

I. The syntactic role of the governor of the acl clause (this is to check Vella's (1831: 225) and Kalmár and Agius' (1983: 338) observations regarding VS order in relative clauses modifying the object);
II. the structure of the acl clause in terms of dependents and their order;
III. heaviness of the subject of the acl clause; and
IV. length of the acl clause.

Each of these items of information was extracted from the PTG files by means of an XSL transformation (see Appendix B for the files). The syntactic role of the governor of the acl clause was recorded by its relation (nsubj, dobj and so on); for the structure of the acl clause, the dependents of the head were extracted by relation and marked according to their position as either pre- or post-verbal; the heaviness of the subject and the length of the clause were recorded as an integer equaling the sum of their dependent tokens. In this manner, I arrived at a list comprising 47 features:

> Heaviness, Length, post-advcl, post-advmod, post-cc, post-ccomp, post-compound, post-conj, post-dobj, post-iobj, post-nmod:advmod, post-nmod:obj, post-parataxis, post-punct, post-remnant, post-xcomp, pre-advcl, pre-advmod, pre-aux, pre-case, pre-cop, pre-dobj, pre-mark, pre-neg, pre-nmod:advmod, pre-nmod:obj, pre-parataxis, pre-part, pre-punct, head-nmod:poss, head-dobj, head-nmod:obj, head-nmod:advmod, head-nmod, head-conj, head-nsubjpass, head-nsubj, head-ccomp, head-nmod:agent, head-appos, head-advcl, head-root, head-advmod, head-acl, head-iobj, head-parataxis, head-xcomp

These were encoded in a CSV file in binary terms, i.e. 1 for the presence of a variable, 0 for its absence; the constituent order configuration was also encoded as a binary, 0 for SV, 1 for VS; heaviness and length were encoded as continuous.

The analysis itself was performed in the R software environment using a generalized linear mixed model (GLMM, Baayen 2008: 278-284) which allows the inclusion of random effects; this makes it particularly useful to sciences like linguistics which deal with human beings (cf. Johnson 2009: 247-265). In R, GLMMs are implemented in the library lme4 (Bates et al. 2015) and for this analysis, I modeled the constituent order configurations as the binary dependent variable with the listed features as predictors. The final code for the full model is given below and essentially replicates the list of the 47 features above. The only difference here is that the feature post-cc has been excluded as a predictor because in this data set, it correlated perfectly with post-conj (coordinating conjunctions usually appear together with conjuncts), so the final number of features is 46, plus a random effect per sentence.

```
full.mod.MIX <- glmer(Order ~ Heaviness + Length + post.advcl
+ post.advmod + post.ccomp + post.compound + post.conj
+ post.dobj + post.iobj + post.nmod.advmod + post.nmod.obj
+ post.parataxis + post.punct + post.remnant + post.xcomp
+ pre.advcl + pre.advmod + pre.aux + pre.case + pre.cop + pre.dobj
+ pre.mark + pre.neg + pre.nmod.advmod  + pre.nmod.obj + pre.parataxis
+ pre.part + pre.punct + head.nmod.poss + head.dobj + head.nmod.obj
+ head.nmod.advmod + head.nmod + head.conj + head.nsubjpass + head.nsubj
+ head.ccomp + head.nmod.agent + head.appos + head.advcl + head.root
+ head.advmod + head.acl + head.iobj + head.parataxis + head.xcomp
+ (1|Sentence),
data=acl.clauses, family="binomial")
```

As the primary test of model fit, I used the function somers2().[4] This function compares actual results (i.e. the values of the dependent variable from the data set) with predictions by the model and produces two measures on the 0-1 scale, the concordance index C and Somer's Dxy rank correlation, where 1 indicates perfect fit. With C at 0.97 and Dxy at 0.94, the fit of the model was deemed excellent (cf. Baayen 2008: 281) and thus the model itself suitable for further analysis, providing requisite caution is exercised in its interpretation.

The first step in the analysis entailed reducing the model by removing the predictors representing the head of the acl clause. The resulting reduced model was then compared to the original model using the anova() function to perform a chi-squared test. With the p-value obtained from the test at 0.833, the null hypothesis stating that the two models are the same cannot be ruled out and thus the models are equivalent in their predictive capabilities; this is only confirmed by the somers2() diagnostic on the reduced model where the C and Dxy measure decreased very slightly to 0.96 and 0.92, respectively. When the same procedure was applied to the features representing the structure of the clause itself, the differences were statistically significant. The conclusion to be drawn here is that contrary to previous analyses, the syntactic role of the governor is not a factor in the order of predicate and subject in acl clauses, while the structure of the clause is.

The next steps in the analysis involved the inspection of the reduced model and its fit, starting with a comparison with a generalized linear model (GLM) where it was found that there is no statistically significant difference between the GLMM and a GLM and thus that the random effect plays no role. The standard battery of tests performed subsequently offered a rather complex picture. On the whole, the reduced models (whether GLMM or GLM) provide an excellent fit. However, only three features were found to be significant as coefficients, whether through the examination of GLMM or the stepwise reduction of the model using the drop1() function for the GLMM model or the stepAIC() function from the MASS library (Venables and Ripley 2002) for the GLM model: heaviness, length and the presence of a post-conj. Heaviness was strongly positively associated with the VS order and the latter two were associated negatively with VS order and thus positively with SV order.

In the final step of the analysis, a model consisting solely of heaviness, length, post-conj and a random effect was first compiled. This model provided a somewhat reasonable fit (C 0.83, Dxy 0.66), however, only subject heaviness and clause length emerged as statistically significant factors predictors of order of subject and predicate in acl clauses. A further model with post-conj removed was found to have the same predictive power (the same anova() chi-squared test as described above comparing the two produced a p-value of 0.25) and so a final GLM model containing only heaviness and length as predictors was compiled with the following properties:

4 bit.ly/2HKGQJA (last consulted on February 28th 2018)

```
> summary(hl.GLM)
Call:
glm(formula = Order ~ Heaviness + Length, family = "binomial",
data = acl.clauses)

Deviance Residuals:
Min       1Q   Median       3Q      Max
-1.6467  -0.9738  -0.4135   1.0044   3.3121

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.32617    0.34466   0.946 0.343973
Heaviness    0.36673    0.09824   3.733 0.000189 ***
Length      -0.15953    0.04001  -3.987 6.68e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 200.64  on 145  degrees of freedom
Residual deviance: 169.23  on 143  degrees of freedom
AIC: 175.23
> somers2(binomial()$linkinv(fitted(hl.GLM)), acl.clauses$Order)
        C         Dxy           n     Missing
0.8250712   0.6501425 146.0000000   0.0000000
```

This, along with the plot in Figure 7.8 (produced using the sjPlot R library, Lüdecke 2017), to some extent confirmed the findings above: the heavier its subject, the more likely the acl clause is VS (recall that in the models, the VS order was encoded as 1); the longer the acl clause, the more likely the clause is SV .

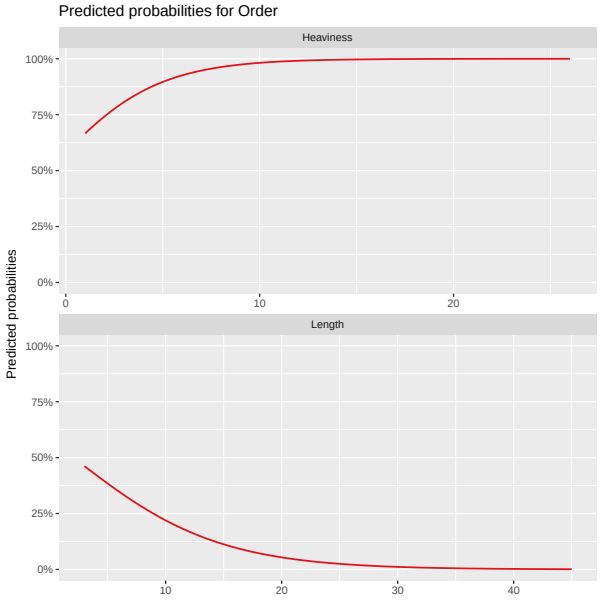


Fig. 7.8: MUDTv1: Subject heaviness and clause length as predictors of the order of S and V in active acl clauses

Consequently, all that can be said with confidence here is that the syntactic role of the governor plays no role in determining the order of subject and predicate in acl clauses and that heaviness of the subject is associated with the VS configuration, while clause length is strongly associated with SV configuration. Based on the examination of the data, one could hypothesize that the structure of the clause itself may influence the configuration as well: for example, post-verbal dobj and adverbials (both advmod and nmod:advmod) appear to favor the placement of the subject before the predicate; a pre-verbal advmod appears to occur more frequently with the VS order. Further work on more data will be required to check these hypotheses and provide the ultimate answer to the question of determinants of constituent order in Maltese acl clauses, taking into account other factors, including semantic ones like the restrictive and non-restrictive distinction (Camilleri and Sadler 2016).

7.3.2.3 Order of predicate and direct object

Figure 7.9 plots the distribution of VO and OV orders where O combines the counts for both dobj and nmod:obj; Table 7.13 provides the same information while adding absolute numbers.

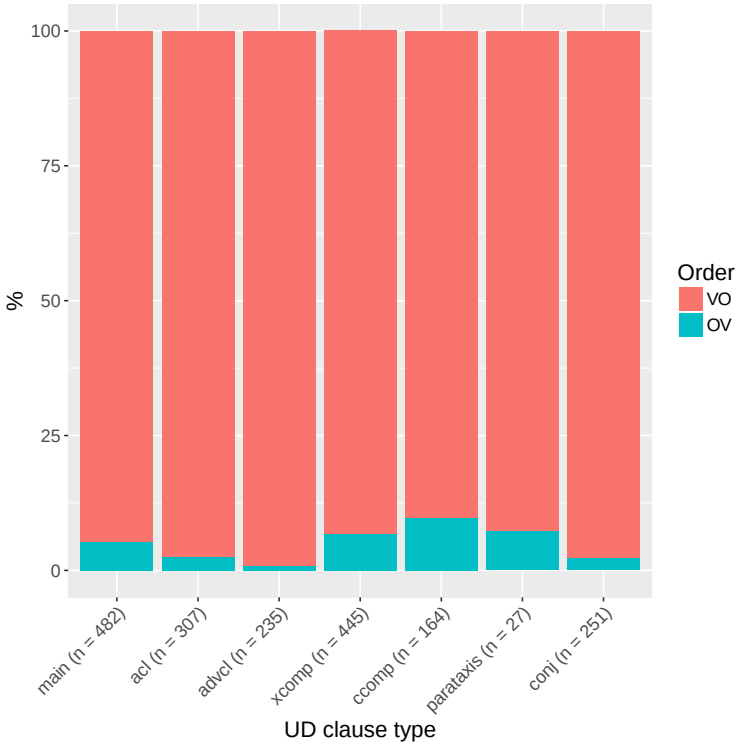


Fig. 7.9: MUDTv1: Order of predicate and direct object in verbal clauses by UD clause type

UD clause type	Order	Count	%
main	VO	456	94.61%
	OV	26	5.39%
acl	VO	299	97.39%
	OV	8	2.61%
advcl	VO	233	99.15%
	OV	2	0.85%
xcomp	VO	415	93.26%
	OV	30	6.74%
ccomp	VO	148	90.24%
	OV	16	9.76%
parataxis	VO	25	92.59%
	OV	2	7.41%
conj	VO	245	97.61%
	OV	6	2.39%
Total	VO	1821	95.29%
	OV	90	4.71%

Tab. 7.13: MUDTv1: Order of predicate and direct object in verbal clauses by UD clause type

One note on the counts for xcomp and parataxis: in both cases, there are instances of dobj which depends on a X_FOR, 1 case for xcomp, 2 for parataxis. Those 3 dobj have been excluded from the table and from futher consideration, as have the 4 dobj in csubj; all 7 are VO.

The numbers above require further elaboration in terms of the categorization of OV clauses, especially with regard to deviation (Research Question 3). As becomes immediately evident upon inspection of these clauses, they fall into two major groups, the first of which is made up of clauses where the object is a PRON_INT or a NOUN with a PRON_INT dependent (i.e. interrogative clauses and ccomp clauses introduced by interrogatives). This appears to confirm Borg and Azzopardi-Alexander's observation that for interrogative pronouns (*x'*, *xi* and *min*), their default position is before the verb (Borg and Azzopardi-Alexander 1997: 210). In fact, for *x'* and *xi*, a special form exists which is used if the interrogative pronoun must follow the verb (Borg and Azzopardi-Alexander 1997: 11, 210); there is only a single instance of *xiex* in MUDTv1 and even then it appears in its other role, questioning "a noun in a prepositional phrase" (Borg and Azzopardi-Alexander 1997: 16). The same is then true of questioned objects (Borg and Azzopardi-Alexander 1997: 12-13) introduced by *xi* or *liema*. In both these cases, we are therefore dealing with constructions where OV is the dominant order (cf. (Borg and Azzopardi-Alexander 1997: 20).

Consequently, it is only the other major group of OV clauses that constitute the actual deviation from the default VO in Maltese. And while a detailed analysis of said deviation is outside the scope of this work, I have nevertheless performed a quick one, for the purposes of checking previous analyses of this phenomenon in terms of its role

in information structure (Fabri 1993, Borg and Azzopardi-Alexander 1997, Borg and Azzopardi-Alexander 2009) and answering Research Question 3. For this analysis, I used the concepts of topic and focus defined in terms of their role in the discourse; in other words, the term "topic" is used for contextually anchored entities (i.e. mentioned previously, whether explicitly or implicitly) and the term "focus" is then used for entities introduced into the discourse. Additionally, I took into account the well-established role of clitics (Fabri 1993: 144-146) and finally, I singled out all nmod:obj preceding their predicate for a separate analysis due to the fact that in this situation, all such nmod:obj are prepositional phrases. Table 7.14 contains the results of the analysis.

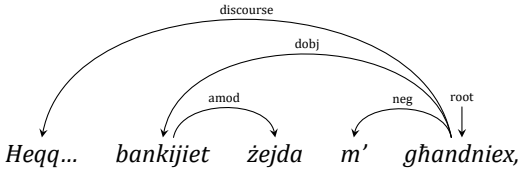
OV clause type	Count	%
Interrogative pronoun	29	32.22%
Questioned noun phrase	6	6.67%
Topic with clitic	44	48.89%
Topic without clitic	4	4.44%
Focus with clitic	1	1.11%
Focus without clitic	3	3.33%
nmod:obj	3	3.33%
Total dominant OV	35	38.89%
Total deviant OV	55	61.11%

Tab. 7.14: MUDTv1: Types of OV clauses

As evident from this data, topicalization of the object is indeed the primary function of the deviant OV order; however (as observed in Fabri and Borg 2002), it is not the only one. Nor is the role of clitics solely reserved – and a necessary condition – for topics, as Fabri (1993: 145-146) and Fabri and Borg (2002: 362) argue. Some of this may be due to structural (morphological) limitations, as two of the topics without a clitic on the verb involve the VERB_PSEU *għand-* "to have" which cannot take direct object clitics, as in (6) where the (nota bene indefinite) direct object *bankijiet* "benches" features in an answer to a request for benches. The other two, however, have no such limitation which underscores the fact that a full account of the role of encliticization in topicalization is still a desideratum.

(6) *Heqq... bankijiet żejda m' għandniex,*
well... bench-PL additional-PL NEG we have-NEG,

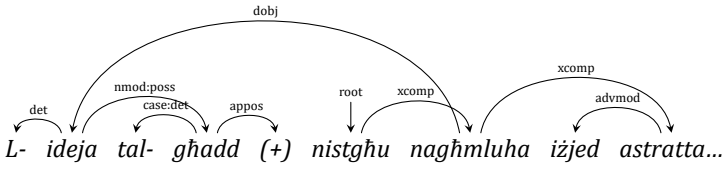
'Well... We don't have extra benches,'



[MUDTv1: 49_03F09]

The same is by and large true of OV clauses where the object is in focus. Such clauses should not feature an encliticized verb (Fabri 1993: 145-146) and yet, they do (7):

(7) *L- ideja tal- għadd (+) nistgħu nagħmluha iżjed astratta...*
DEF idea GEN-DEF addition (+) we can we make her more abstract-F...
'We can make the idea of addition (+) more abstract...'



[MUDTv1: 55_02N11]

In this example (the 8th sentence in the text), the dobj introduces entities that are completely new to the discourse, both in terms of concepts (semantics), as well as in terms of the actual realization (the lexeme), without any contrast and, this being a written text, no stress (cf. Fabri and Borg 2002).

The primary takeaway from this analysis, however, involves topicalized objects: counting both those with clitics and those without, we arrive at the total number of 48 out of 1911 analyzed direct objects, for a rate of 2.5%. This contradicts the description of such constructions as "a wide spread characteristic of Maltese" (Borg and Azzopardi-Alexander 1997: 126).

And finally, we turn to OV clauses where the O is a nmod:obj. As the numbers in Table 7.13 combined the counts for dobj and nmod:obj, first, an overview of the distribution of VO/OV configurations for each type of object separately is provided in Table 7.15.

MUDTv1 object type	Order	Count	%
dobj	VO	1525	94.72%
	OV	85	5.28%
nmod:obj	VO	296	98.34%
	OV	5	1.66%

Tab. 7.15: MUDTv1: dobj and nmod:obj

As evident from Table 7.14, 3 of those 5 nmod:obj are OV deviations of the default VO order and all 3 are instances of object in focus. The remaining two OV nmod:obj then occur in subordinate clauses and have interrogative pronouns as heads.

7.3.2.4 Order of predicate and indirect object

In typological analyses of constituent order, the indirect object is often omitted from consideration (cf. Comrie 1989: 89) and MUDTv1 data provides some justification for this in quantitative terms: there are only 77 cases of iobj in MUDTv1, as compared to 1911 for dobj and nmod:obj combined. Figure 7.10 plots the distribution of the order of indirect object (I) and the predicate (V) across UD clause types, Table 7.16 provides the same information in absolute as well as relative numbers.

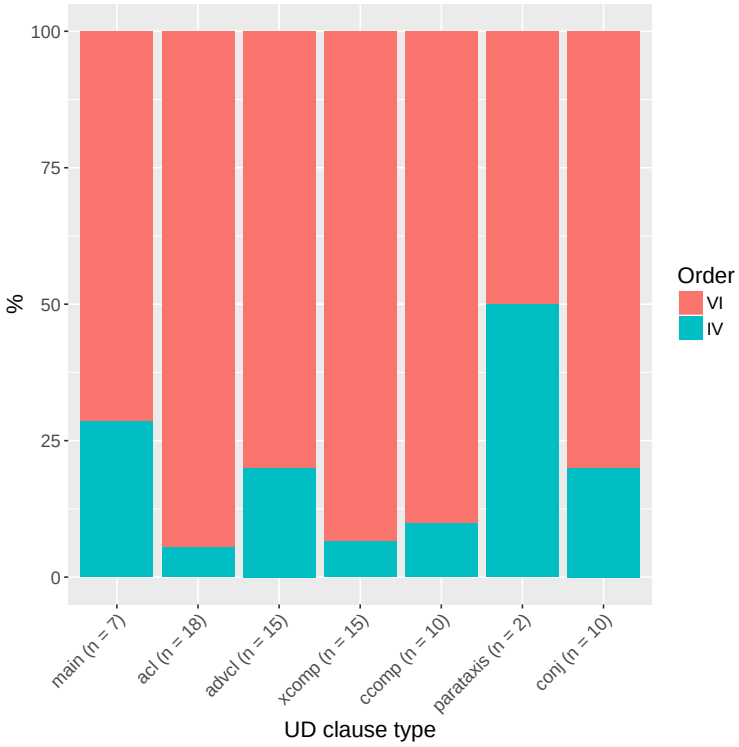


Fig. 7.10: MUDTv1: Order of predicate and indirect object in verbal clauses by UD clause type

UD clause type	Order	Count	%
main	VI	5	71.43%
	IV	2	28.57%
acl	VI	17	94.44%
	IV	1	5.56%
advcl	VI	12	80.00%
	IV	3	20.00%
xcomp	VI	14	93.33%
	IV	1	6.67%
ccomp	VI	9	90.00%
	IV	1	10.00%
parataxis	VI	1	50.00%
	IV	1	50.00%
conj	VI	8	80.00%
	IV	2	20.00%
Total	VI	66	85.71%
	IV	11	14.29%

Tab. 7.16: MUDTv1: Order of predicate and indirect object in verbal clauses by UD clause type

The same analysis of OV deviation as performed for direct object above could be attempted here, but regarding the low number of IV clauses, it will hardly be conclusive. And so I tentatively note that of the 11 IV clauses, 2 are actually cases of dominant IV involving the verb *sejjaħ* "to call, to give a name", 3 are impersonal (the so-called ethical dative), 4 involve focus (including contrastive focus) and only 2 involve actual topicalization, for a share of 2.6%.

For completeness' sake, I also include here the statistics for the order of ditransitive clauses, i.e. clauses featuring both a direct object (dobj or nmod:obj) and an indirect object (iobj). There are only 36 such clauses in MUDTv1, all of them active verbal clauses; the distribution of configurations of predicate (V), direct object (O) and indirect object (I) is plotted in Figure 7.11 below while Table 7.11 contains the absolute numbers. As evident from these numbers, the VOI order is the dominant one in Maltese.

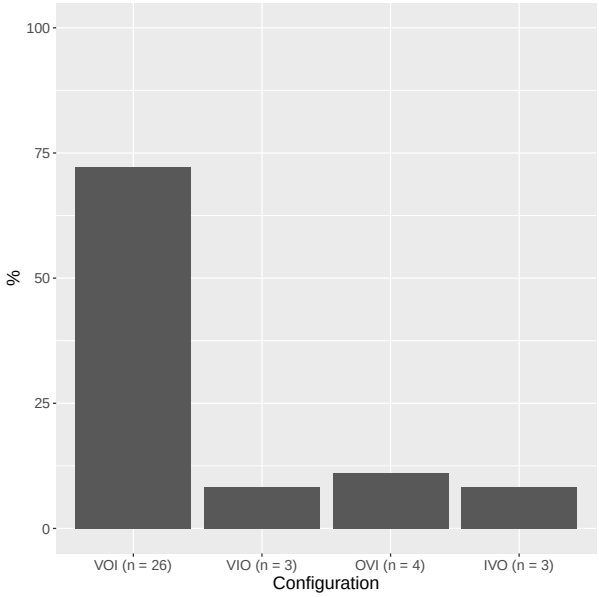


Fig. 7.11: MUDTv1: Order of predicate, direct object and indirect object

Configuration	Count	%
VOI	26	72.22%
VIO	3	8.33%
OVI	4	11.11%
IVO	3	8.33%
Total	36	100%

Tab. 7.17: MUDTv1: Order of predicate, direct object and indirect object

7.3.2.5 Order of passive subject and predicate

7.3.2.5.1 General

Figure 7.12 plots the distribution of orders of passive subject (nsubjpass) and predicate in passive clauses; Table 7.18 provides the same information while adding absolute numbers.

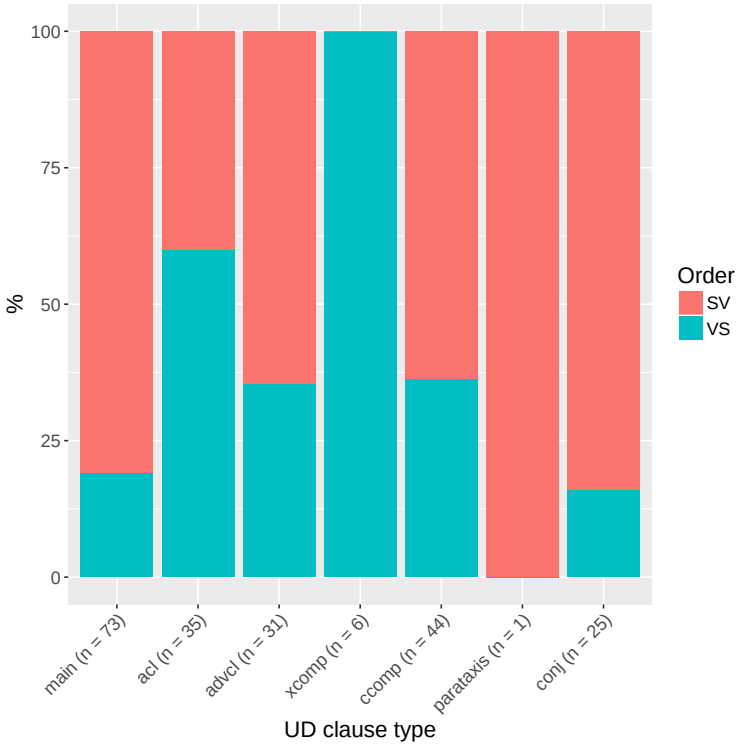


Fig. 7.12: MUDTv1: Order of subject and predicate in passive clauses by UD clause type

UD clause type	Order	Count	%
main	SV	59	80.82%
	VS	14	19.18%
acl	SV	14	40.00%
	VS	21	60.00%
advcl	SV	20	64.52%
	VS	11	35.48%
xcomp	SV	0	0.00%
	VS	6	100.00%
ccomp	SV	28	63.64%
	VS	16	36.36%
parataxis	SV	1	100.00%
	VS	0	0.00%
conj	SV	21	84.00%
	VS	4	16.00%
Total	SV	143	66.51%
	VS	72	33.49%

Tab. 7.18: MUDTv1: Order of subject and predicate in passive clauses by UD clause type

As evident from this data, in some aspects, passive clauses behave much like active clauses do, in that passive xcomp shows a clear preference for VS and no dominant order can be established for acl (although this time, the ratio is closer to Dryer's 2:1). In contrast to verbal clauses, however, there are two types of passive clauses that barely skirt the lower limits of Dryer's ratio: advcl and ccomp. This brings us to an important practical question: what to do when the ratio of two configurations is almost 2:1, but not quite? With such small number of clauses, it is quite possible that the difference is due to a single clause, as with ccomp where moving one clause from the VS row to the SV row would result in a 68%-32% split and thus a clear classification of passive ccomp as SV. Since in such cases, the ratio of SV and VS could be due to chance, a test of statistical significance of these differences should be applied to determine whether they are real.

The solution I adopted here is based on the fact that this is quite obviously the binomial problem: given an event with two outcomes (in our case, SV and VS) and *N* trials (the total number of clauses of a particular type with nsubj), what is the probability that the obtained outcome *k* is different from the expected one (i.e. 50/50 distribution of both configurations)? As standard binomial probability calculation is somewhat problematic in this context (Wallis 2013), I employed the following method based on confidence intervals (Milička 2014): first, I calculated the 95-percentile confidence intervals for the probability of VS order as recorded in Table 7.18 for both bordeline cases (advcl and ccomp), as well as for acl just in case and for main passive clauses to serve as a con-

trol of sorts. To do this, I used the function binconf() from the R Hmisc package:[5] this function takes the obtained outcomes (in our case, the number of clauses with VS order, but the SV order would serve just as well) and the total number of trials (*N*) and using the Wilson score test (see Wallis 2013: 183-189 for the reasoning behind its use), calculates and returns the probability of the supplied outcome (PointEst), together with the lower and upper bounds of its confidence interval:

```
> binconf(21,35) #passive acl clauses
PointEst Lower     Upper
0.6   0.4357271 0.7444927
```

The two latter figures returned by binconf() I then supplied to the function rbinom()[6] to generate a vector of 100 random samples out of the total number of clauses (*N* supplied as the size argument in the function below), 50[7] for each of the extremes of the confidence interval.

```
> acl <- c(rbinom(n = 100, size = 35, prob = binconf(21,35)[,2]), #lower
rbinom(n = 100, size = 35, prob = binconf(21,35)[,3])) #upper
> head(acl)
[1] 20 16 18 18 22 17
```

In this manner, probabilities in the confidence interval are turned to absolute numbers where each item in each of the four vectors lies within the 95% confidence interval. These vectors can then be used for comparison with real-life data as follows: recall that to determine dominant constituent order, one of the configuration pairs would have to be represented in less than 33% of clauses. As the random samples obtained in the previous step reflect the distribution of VS order (which, based on actual MUDTv1 numbers, appears to be the non-dominant order), all that needs to be done is to determine whether the entire spread of values is lower than the 33% threshold for each clause type in question (Table 7.19).

Clause type	Total (N)	33%
acl	35	10.39
advcl	31	10.33
ccomp	44	14.67
main (control)	73	24.33

Tab. 7.19: MUDTv1: 33% threshold for determining dominant constituent order in borderline cases

5 bit.ly/2ESVTD5 (last consulted on February 28th, 2018)

6 bit.ly/2CIsqGf (last consulted on February 28th, 2018)

7 1 would technically have been enough, but why not go the extra mile.

For the actual comparison, I opted for a visual inspection and plotted the sampling data in boxplots with the 33% threshold expressed as the respective absolute number as a horizontal line (Figure 7.13). As the cutoff, I established the following rule: for a particular clause type to pass, the 1st and 3rd quartiles (the box itself) would have to be below their respective horizontal lines.

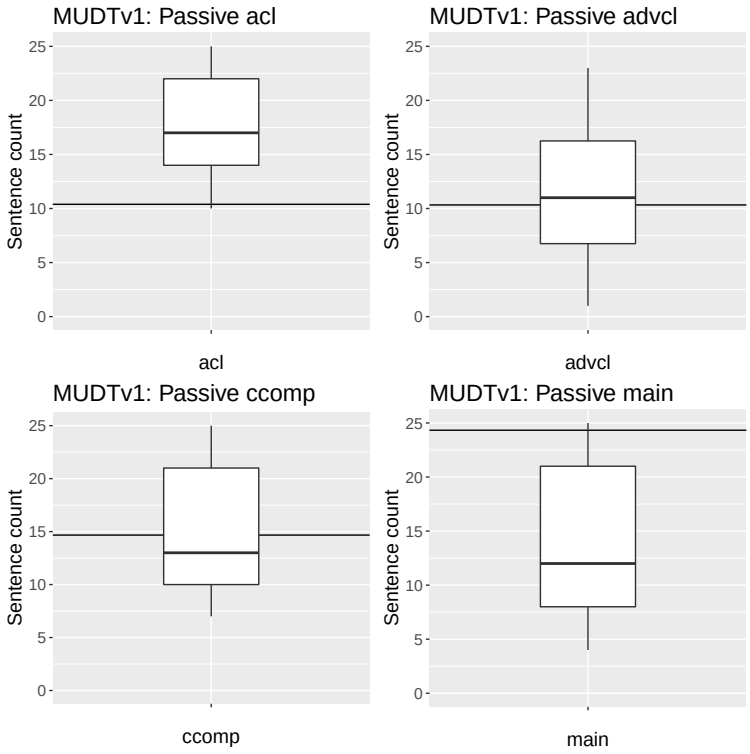


Fig. 7.13: MUDTv1: Sampled VS order in borderline cases with main as control

Both borderline cases fail this test, advcl thoroughly so, as not even the median line in the boxplot drops beneath the threshold. The same is true of acl, but this is hardly unexpected; nor is the fact that main clauses pass, whiskers and all. This leads me to the conclusion that for all the three clauses in question, passive acl, passive advcl and passive ccomp, no dominant order of subject and predicate can be established.

This, in turn, necessitates an explanation as to the factors that cause this variation. Same analysis as the one in active acl was attempted for each of the three clause types; first per clause type, which it was inconclusive due to lack of data, then across all three clause types. In this analysis, subject heaviness and clause length once again emerged

as significant predictors of VS and SV orders, respectively. Any definite conclusion, however, will require more data than is available in MUDTv1.

7.3.2.5.2 VS in passive xcomp

Like their active counterparts, passive xcomp clauses seem to exhibit VS as the dominant constituent order; Table 7.20 below summarizes the data for the distribution of orders of predicate and nsubjpass in passive xcomp clauses.

UD clause type	Order	Count	%
Passive xcomp	SV	0	0.00%
	VS	6	100.00%

Tab. 7.20: MUDTv1: Dominant VS in passive xcomp

In this case, the same considerations as those for active xcomp clauses apply (see section 7.3.2.2.2). The clauses that are the heads of the passive xcomp clauses in question and thus the actual governors of their nsubjpass are listed in Table 7.21.

Head	Count	%
main	1	16.67%
acl	1	16.67%
advcl	1	16.67%
ccomp	3	50.00%
Total	6	100.00%

Tab. 7.21: MUDTv1: Head of passive xcomp

Unlike with active xcomp clauses, there is no passive xcomp clause proper with a nominal subject.

7.3.2.6 Order of predicate and passive agent

Figure 7.14 plots the distribution of orders of predicate (V) and agent (A) in passive clauses across UD clause types, Table 7.22 provides the same data in absolute and relative numbers.

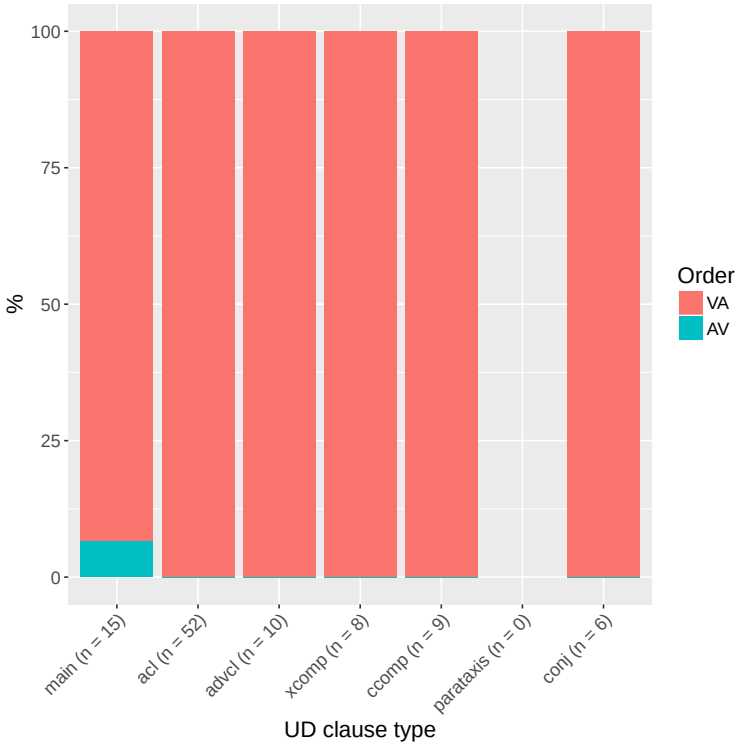


Fig. 7.14: MUDTv1: Order of passive agent and predicate in passive clauses by UD clause type

UD clause type	Order	Count	%
main	VA	14	93.33%
	AV	1	6.67%
acl	VA	52	100.00%
	AV	0	0.00%
advcl	VA	10	100.00%
	AV	0	0.00%
xcomp	VA	8	100.00%
	AV	0	0.00%
ccomp	VA	9	100.00%
	AV	0	0.00%
parataxis	VA	0	NA
	AV	0	NA
conj	VA	6	100.00%
	AV	0	0.00%
Total	VA	99	99.00%
	AV	1	1.00%

Tab. 7.22: MUDTv1: Order of passive agent and predicate in passive clauses by UD clause type

The most interesting thing to note here are the absolute numbers: the AV stack in the "main" clause column represents a single AV construction in MUDTv1 out of 100. The entire sentence in question with the A in question underlined is provided below as (8).

(8) *Għalkemm il- ġenituri offrew li jagħtu l- organi*
although DEF parent-PL they offered COMP they give DEF organ-PL
tagħha, mill- isptar kienu nfurmati li peress li
her, from-DEF hospital they were informed-PL COMP because COMP
kellha infezzjoni ma riedux jieħdu riskju fuq ħaddieħor.
she had infection NEG they wanted-NEG they take risk on someone else.
'Although the parents offered to donate her organs, they were informed by the hospital that because she had infection, they didn't want to expose someone else to a risk.'

[MUDTv1: 05_05J01]

7.3.3 Copular clauses

7.3.3.1 Overview

Figure 7.15 plots the distribution of configurations in copular clauses across all UD clause types; Table 7.23 provides the same information including absolute numbers, excluding a solitary instance of a copular csubj.

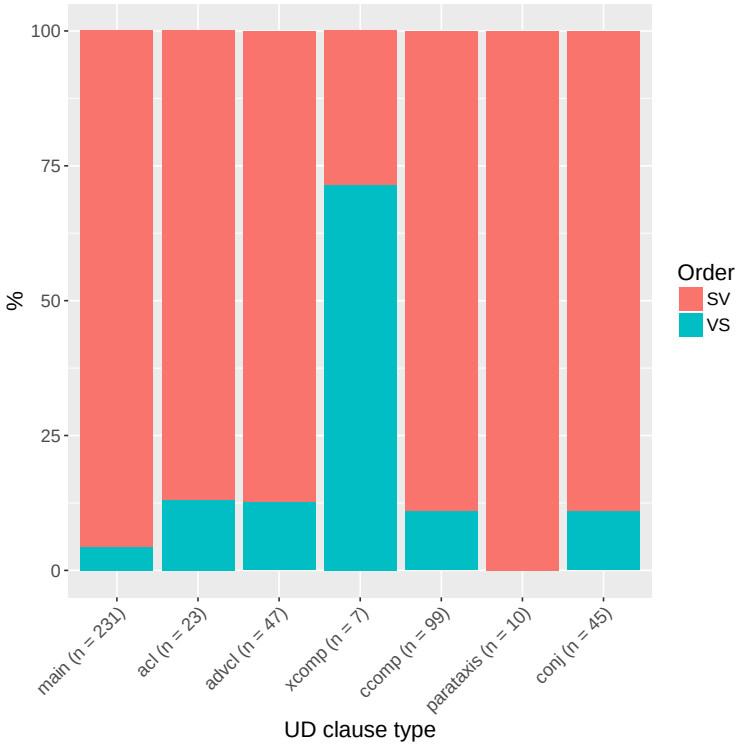


Fig. 7.15: MUDTv1: Order of subject and predicate in copular clauses by UD clause type

UD clause type	Order	Count	%
main	SV	221	95.67%
	VS	10	4.33%
acl	SV	20	86.96%
	VS	3	13.04%
advcl	SV	41	87.23%
	VS	6	12.77%
xcomp	SV	2	28.57%
	VS	5	71.43%
ccomp	SV	88	88.89%
	VS	11	11.11%
parataxis	SV	10	100.00%
	VS	0	0.00%
conj	SV	40	88.89%
	VS	5	11.11%
Total	SV	422	91.34%
	VS	40	8.66%

Tab. 7.23: MUDTv1: Order of subject and predicate in copular clauses by UD clause type

7.3.3.2 VS in copular xcomp

The dominant VS order in xcomp copular clauses (summarized in Table 7.24) requires an explanation different from the one given for same phenomenon in verbal clauses.

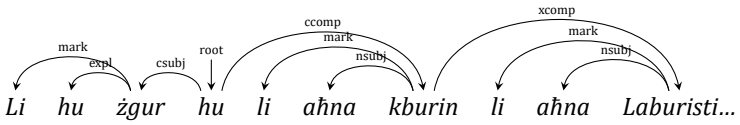
UD clause type	Order	Count	%
xcomp	SV	2	28.57%
	VS	5	71.43%

Tab. 7.24: MUDTv1: Dominant VS in copular xcomp

This is largely due to the nature of copular clauses: unlike verbal clauses where the subject can be expressed in verbal affixes, copular clauses (with the exception of those featuring KIEN) require an overt subject and ipso facto, they cannot comply with the primary definition of an xcomp – a complement clause which inherits its subject from a higher clause. A case could be made for annotating such clauses as ccomp as per the rule "overt subject equals ccomp" established in Chapter 6, section 6.4.4.4.4, but as the same section explains, the place of the nsubj in (9) could be taken by KIEN, in which case there would be no doubt as to the status of the clause as an xcomp.

(9) *Li hu żgur hu li aħna kburin li aħna Laburisti...*
COMP he certain he COMP we proud-PL COMP we Laborist-PL...

'What is certain is that we are proud that we are Laborists...'

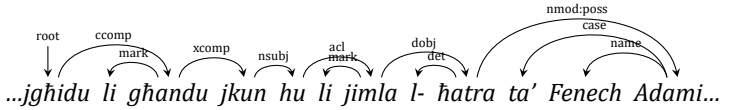


[MUDTv1: 22_02J03]

The copular xcomp clause in (9) is therefore a proper xcomp and together with another clause which exhibits an equivalent structure (an xcomp modifying an adjectival copular predicate), they make up the 2 copular xcomp clauses with SV in MUDTv1.

As for the variation, the five copular xcomp clauses that exhibit VS order fall into three categories: in the first (comprising two clauses, both found in file 04_04J01 and in two consecutive sentences, 4 and 5), the copular clause (the verb KIEN) is the final link in a verbal chain (both consisting of VERB_PSEU *għand-* + KIEN) and its subject is modified by an acl (10).

(10) *...jgħidu li għandu jkun hu li jimla l- ħatra ta' Fenech Adami...*
they say COMP he has he is he COMP he fills DEF post GEN Fenech Adami...
'... they say it has to be him who fills the vacancy left by Fenech Adami...'



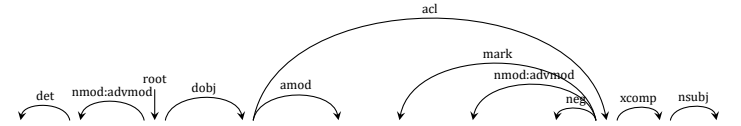
[MUDTv1: 04_04J01]

In this case, the same explanation regarding the nature of verbal chains can be offered as the one for VS verbal xcomp clauses (see section 7.3.2.2.2 above), that we are looking here at a deviation from the default order of the verbal chain and its subject where the verbal chain fulfills the syntactic role assigned to its first member, whatever that may be. And so as the numbers for VS verbal xcomp would be distributed across all other clause types (see Table 7.8), (10) and the other example in this type of VS copular xcomp clauses will be added to the general description of constituent order configurations (Table 7.23, both under ccomp) and subtracted from the count of VS copular xcomp clauses.

The second type of VS copular xcomp clauses in MUDTv1 consists of a solitary case of a relatively straightforward locative copular clause modifying the verb *af* "to know" (11). The predicate here is a locative interrogative pronoun which appears at the beginning of a clause both questions (Borg and Azzopardi-Alexander 1997: 8, 24), adjectival clauses (Borg and Azzopardi-Alexander 1997: 212); the same seems to apply to com-

plement clauses, at least as represented in MUDTv1, where in all such clauses, this is the case with *fejn*.

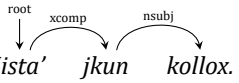
(11) *...xi drabi joħloq karattri ċentrali li magħhom ma*
...some time-PL he creates character-PL central-PL COMP with them NEG
tafx fejn int...
you know-NEG where you...
'...sometimes he creates central characters with whom you don't know where you are ...'



[MUDTv1: 50_01N10]

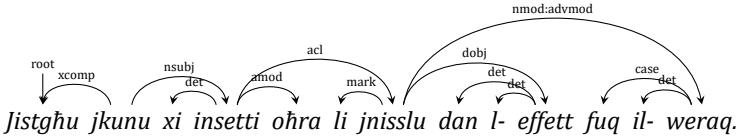
The third and final type of VS copular xcomp clauses in MUDTv1 comprises two clauses and it also the most interesting in its properties: both these clauses are the last links in a verbal chain which begins with the verb *seta' "*can, to be able to" (12, 13).

(12) *Jista' jkun kollox.*
he can he is everything.
'Could be anything.'



[MUDTv1: 18_05J02]

(13) *Jistgħu jkunu xi insetti oħra li jnisslu dan l-*
they can they are some insect-PL other-PL COMP they cause this.M DEF
effett fuq il- weraq.
effect on DEF leaf.PL
'There can be some other insects which have this effect on the leaves.'



[MUDTv1: 56_03N11]

The interesting part here is that while these are copular clauses, they do not express any of the semantic relationship typically associated with copular clauses (Dixon 2010: 159) like identity (which is what (9) denotes), attribution or location (11). Instead, they appear to denote potential existence or presence of their subject: (12) comes after a list of explanations given for an event, sarcastically noting that with the number of options offered, the list could containg anything. In (13), the clause follows an explanation of how a particular type of insect can damage leaves of a particular plant, and it notes that there could be other types of insects which can cause the same damage. Both these clauses therefore introduce new entities into the discourse, unlike (10) where the nsubj has a referent already introduced (and extensively discussed) in the opinion piece in question. Consequently, both (12) and (13) would best be described as existential (of the presentational type) and thus thetic (cf. Sasse 1987, see also Chapter 3, section 3.6). Such clauses cross-linguistically consistently favor the predicate-subject order (Givón 2001b: 257, see also Sasse 1987: 540 for Egyptian Arabic and the discussion of existential clauses below).

Having excluded (10) and its sister clause, the distribution of constituent order configurations in copular xcomp clauses and their subtypes can be summarized as follows (7.25):

Construction type	Order	Count	%
xcomp	SV	2	40.00%
xcomp	VS	1	20.00%
Existential xcomp	VS	2	40.00%
Total		5	100%

Tab. 7.25: MUDTv1: Classification of constituent order variation in copular xcomp

It would therefore appear that upon closer inspection, one must conclude that no dominant order can be established in copular xcomp clauses in MUDTv1.

7.3.4 Existential clauses

Figure 7.16 plots the distribution of configurations in copular clauses across all UD clause types; Table 7.26 provides the same information including absolute numbers.

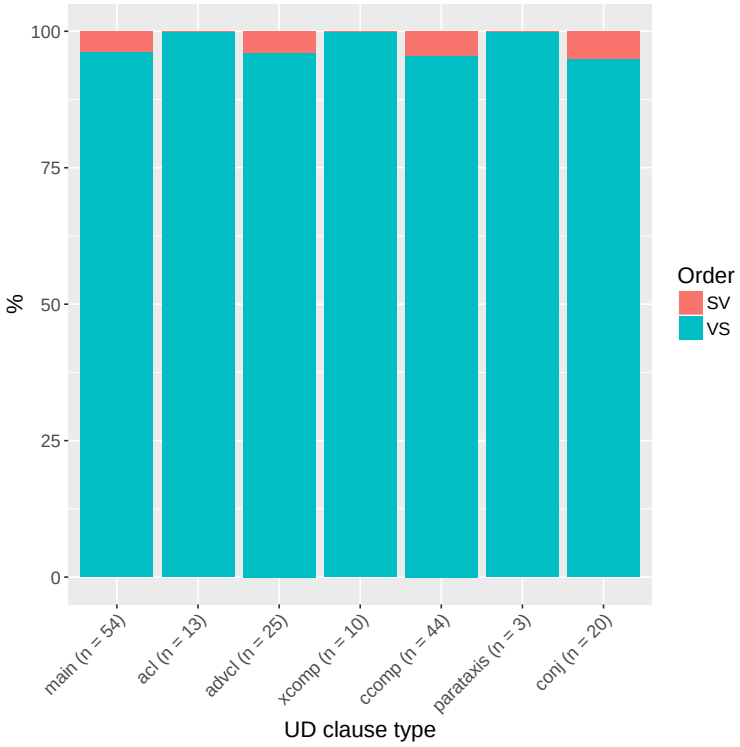


Fig. 7.16: MUDTv1: Order of subject and predicate in existential clauses by UD clause type

UD clause type	Order	Count	%
main	SV	2	3.70%
	VS	52	96.30%
acl	SV	0	0.00%
	VS	13	100.00%
advcl	SV	1	4.00%
	VS	24	96.00%
xcomp	SV	0	0.00%
	VS	10	100.00%
ccomp	SV	2	4.55%
	VS	42	95.45%
parataxis	SV	0	0.00%
	VS	3	100.00%
conj	SV	1	5.00%
	VS	19	95.00%
Total	SV	6	3.55%
	VS	163	96.45%

Tab. 7.26: MUDTv1: Order of subject and predicate in existential clauses by UD clause type

As evident from the data above, existential clauses stand out as the only type of clause defined by root which exhibits VS as the dominant constituent order across all UD clause types (see also Kalmár and Agius 1983: 343-344). This is hardly surprising: as has been noted on many occasions (Givón 2001b: 257, Bentley 2015a: 1), VS appears to be preferred order in existential clauses even in languages which otherwise show clear preference for SV. There are various explanations for this: some argue that this is due to the (inherent) theticity of existential clauses (Sasse 1995: 14-15), others also analyze them in terms of information structure, but argue for some form of topic-comment structure (McNally 2011: 1833). In contrast, Givón (2001b) questions the role of information structure in existentials and argues that the "seeming VS order" in existential clauses is "the consequence of the diachronic pathway of grammaticalization" (Givón 2001b: 259).

Considering the disagreement on the subject, the ultimate answer to the question why that is will have to be answered elsewhere and possibly in comparison with other languages. For the purposes of this chapter, it suffices to conclude that Maltese is one of those languages where VS order is the dominant one in existential clauses and, at the same time, that existential clauses are the only clause type (defined by its root) in MUDTv1 which exhibit this particular configuration as the dominant one.

7.3.5 Constituent order across text types

The data presented in the previous sections would indicate that Maltese as represented in MUDTv1 is a SV/VO language with the deviations representing 23.7% of the clauses

examined for VS and 4.7% for OV. In the final step in this analysis, I provide a breakdown of the deviant orders by text type (Table 7.27).

Text type	% of VS	% of OV
newspaper	25.5 %	2.8%
quasi-spoken	29.7 %	4.2%
fiction	23.9 %	4.8%
non-fiction	15.5 %	4.9%

Tab. 7.27: MUDTv1: Ratio of VS and OV across text types

This data underscores the differences between text types highlighted in section 7.2.1 where again the text type newspaper stands out, this time as the text type with the lowest rate of OV (2.8%). Surprisingly, it is also the non-fiction text type that is the odd one out here, with the share of VS much lower than the average across MUDTv1.

The former defies an easy explanation, but is an important fact in and of itself. As for the latter, comparison across UD clause types (Figure 7.17) would suggest that this is largely due to the combined effect of parataxis and advcl clauses (see the highlighted parts of the graph): non-fiction is the only text type where SV parataxis clauses predominate (albeit only 14 to 10 in absolute numbers) which is hardly surprising considering that this is the only text-type that does not prominently feature reported speech parataxis clauses. This combines with advcl where in non-fiction, the ratio of VS advcl is much lower (15.5%) than in other text types (28%-39%).

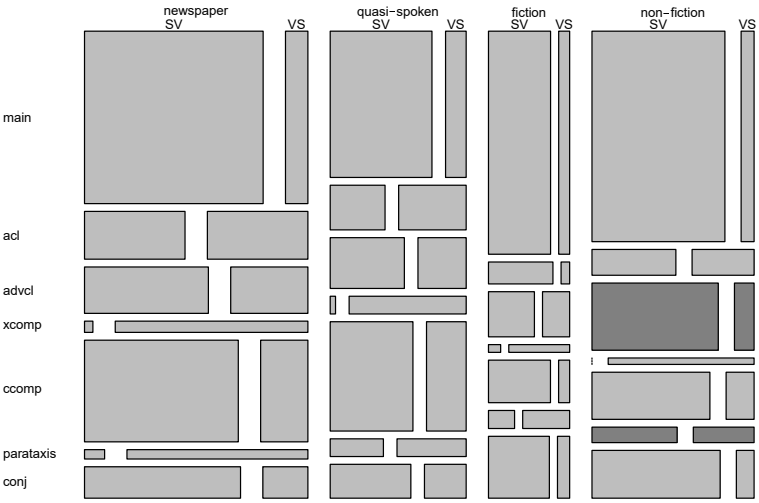


Fig. 7.17: MUDTv1: Order of subject and predicate by text type and UD clause type

7.4 A brief comparison, or: a két fadatbázis regénye

7.4.1 Introduction

On a number of occasions, Maltese has been described as a discourse-configurational language, either explicitly (Fabri and Borg 2002 and Borg and Fabri 2016, both citing Kiss 1995a), or implicitly: Fabri 2010 and Fabri and Borg 2017 describe Maltese as "a topic-oriented language" (Fabri 2010: 793, see also the almost identical phrasing in Fabri and Borg 2017: 83). I take this to be a synonym of "topic-prominent language" (Kiss 1995b: 4-5), a term which sensu stricto designates a subset of languages falling under the "discourse-configurational" umbrella, the so-called type A discourse-configurational languages, where any topicalized constituent can assume the preverbal position typically reserved for the subject (Kiss 1995b: 6-7). In type B discourse-configurational languages, focus-prominent languages, the same is true of focus (Kiss 1995b: 15-24); discourse-configurational languages can be type A, type B or both, modulo interaction between topic and focus and inter-language variation. Those works that describe Maltese as discourse-configurational do not elaborate on that particular aspect of this property, but judging from description of focus provided by Fabri 1993 and Fabri and Borg 2002 (see Chapter 3, section 3.8 and 3.10), if Maltese is a discourse-configurational language, it is both type A and type B. This, however, is ultimately irrelevant: Maltese has been described at least twice as discourse-configurational without any elaboration or qualification and it is this description that is the focus of this section.

As noted in Chapter 2, section 2.3.5, the framework-dependent reasoning behind this classification is of not if interest here. What is, however, is the classification itself, i.e. the claim that Maltese is a discourse-configurational language; more specifically, what I want to focus on is the fact that this claim can be (to some extent) tested. The line of thinking that leads me here is the following:

1. Hungarian is considered the paragon of a discourse-configurational language (cf. Kiss 1995a), i.e. a member of a class of languages defined by a shared property involving constituent order.
2. Maltese has also been described as a discourse-configurational language.
3. Ergo, if one were to investigate the distribution of constituent order configurations in both, one would find that it is at the very least quite similar.

One might also expect that in any discourse-configurational language (and thus both Maltese and Hungarian under assumptions 1 and 2 above), the distribution of SV and VS on one hand and VO and OV on the other would be approximately the same, i.e. 50-50 for both pairs. This is, of course, not realistic, as the theory behind the classification of discourse-configurational languages makes clear: the ordering of constituents is not

random,[8] but based on pragmatic (and possibly other) criteria. Additionally, the subject is more likely to be the topic (as there is a "close correspondence between the topic and the grammatical subject", Kiss 1995b: 10) and in any case, there are inter-language differences in how far discourse-configurationality goes. Nevertheless, the hypothesis above stands and with MUDTv1 and a Hungarian UD v2 treebank[9] (Nivre, Agić et al. 2017), there is a way to test it quantitatively.

7.4.2 Data and analysis

For the purposes of quantitative comparison, I imported the Hungarian UD v2 treebank (henceforth: HUUDv2)[10] into the same instance of ANNIS3 where MUDTv1 resides.[11] Using the ANNIS3 interface, I ran the queries I used for the quantitative analysis of MUDTv1 in section 7.3 above (taking into account the changes from UD v1 to UD v2 and the Maltese specific UD relations) on HUUDv2. I then plotted the two sets of numbers against each other, starting with the comparison of Greenbergian six-way classification for both treebanks (Figure 7.18).

8 On the other hand, both Maltese (Fabri 2010: 793) and Hungarian (Puskás 2000: 41) have been described as having "free word order", so a case could be made that the constituent order in such languages is indeed random.

9 Hence the subtitle of this section, best translated as "a tale of two treebanks". Having failed to find a commonly used (or indeed any) Hungarian translation of "treebank", I came up with my own, a portmanteau of *fa* "tree" and *adatbázis* "database".

10 As noted in Chapter 6, section 6.5.2, MUDTv1 and HUUDv2 are very similar in size: 2047 sentences in MUDTv1, 1800 in HUUDv2; 44,162 tokens in MUDTv1, 42,032 in HUUDv2.

11 bulbul.sk/annis-gui-3.4.4

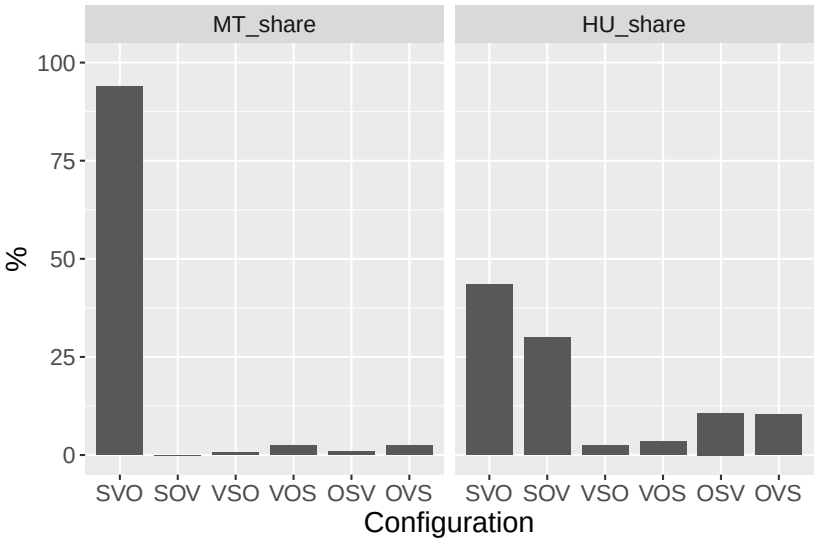


Fig. 7.18: MUDTv1 vs HUUDv2: Constituent order – Greenbergian classification

These two, needless to say, are not the same or even similar. If one were inclined to employ Dryer's 2:1 criteria (Dryer 2013a) here, two different classifications would have to be employed: Maltese (as represented in MUDTv1) as a language with SVO dominant constituent order; Hungarian (as represented in HUUDv2) as a language with no dominant constituent order.

As in this work, Dryer's binary typology is the central paradigm of constituent order typology, let us proceed to that analysis, first plotting a comparison between the distribution of SV and VS orders in both treebanks across UD clause types (Figure 7.19). For the purposes of this comparison, I removed the UD clause type xcomp from the Maltese data for reasons explained in section 7.3.2.2.2 above; in any case, this relation is used for two very different phenomena in the two languages: in Hungarian, these are typically infinitives which do not take subjects.

Fig. 7.19: MUDTv1 vs HUUDv2: Order of subject and predicate by UD clause type

Here the picture is a little more complicated, so in order to make sense of it and provide a test of statistical significance, I plotted a comparison of the means of the two data sets (i.e. the ratios of SV and VS order per clause type) with confidence intervals (Milička 2014) obtained from bootstrap resampling calculated using the R library boot (Canty and Ripley 2017, Davison and Hinkley 1997) as Figure 7.20.

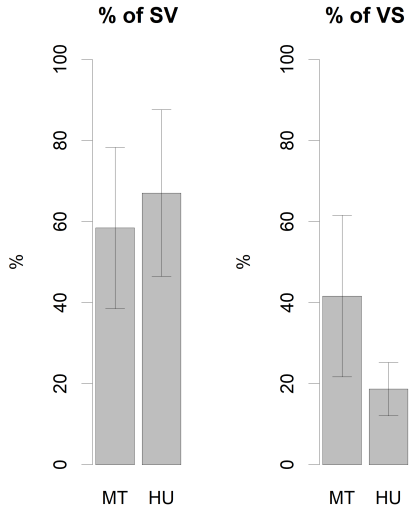


Fig. 7.20: MUDTv1 vs HUUDv2: SV vs VS order

This comparison shows the first signs that while Maltese (as represented in MUDTv1) and Hungarian (as represented in HUUDv2) can be safely classified as SV in Dryer's binary typology, the two languages do not behave similarly when considered more closely. The probability of SV order occurring is similar in both languages, but this is not true of VS: in Maltese, the average share of VS order across all UD clauses is 42.2%; in Hungarian, it is 25.2%. The difference is largely due to two types of Maltese clauses, acl and parataxis, both of which are somewhat problematic (and parataxis even more so, as discussed above); in contrast, the share of the VS configuration is much more uniform in Hungarian. And so it is especially because of Maltese acl clauses that we cannot conclude that Maltese and Hungarian behave almost identically when it comes to the order of subject and predicate: as noted above (and evident from Figure 7.20), there seem to be no dominant or even preferred order in acl clauses in Maltese. As such, Maltese acl clauses – and only these clauses – are much more flexible in their ordering of subject and predicate than Hungarian ones, which is surprising: one would expect to be the share of the deviant order to be roughly the same across all clause types, if pragmatic factors were the only or the primary determining factor in ordering of the constituents; and in fact, this is what one finds in Hungarian. That the ratio of SV and VS in Hungarian is smaller than 50% is also unsurprising: as Kiss points out (1995b: 10), grammatical subjects are more likely to be topics and thus the subject-predicate order will coincide with topic-predicate order.

A plot of the distribution of VO and OV configurations across UD clause types in both treebanks (Figure 7.21) provides a much clearer picture:

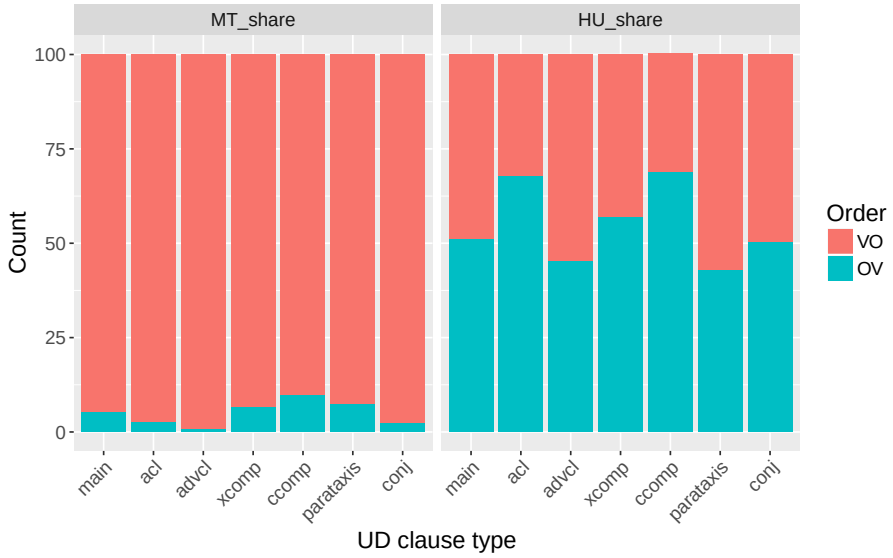


Fig. 7.21: MUDTv1 vs HUUDv2: Order of predicate and object by UD clause type

Just to hammer the point home, the same comparison of the two sets of data with confidence intervals as the one provided for VS and SV configurations is included here as Figure 7.22.

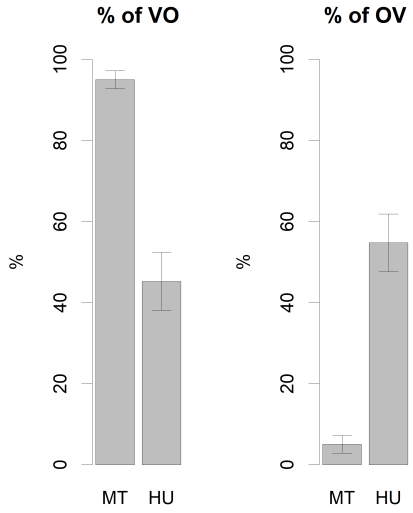


Fig. 7.22: MUDTv1 vs HUUDv2: VO vs OV order

There are two main takeaways here: first, the sharp difference between MUDTv1 and HUUDv2. Secondly, there is the relatively uniform distribution of VO and OV in HUUDv2 which echoes that of SV/VS, but which reaches an average of 50% across all clause types. This not only indicates that Hungarian (as represented in HUUDv2) cannot be classified as either a VO or an OV language, but it also conforms to the naive expectation regarding constituent order variation in discourse-configurational languages expressed above: the probability of VO and OV configurations occurring is roughly the same not only across the board, but also across all clause types. If one were interested in the theory, then one could argue that this is perfectly consistent with it: the roughly 50-50 distribution of VO and OV is what one would expect if the position of the object were only determined by information structure considerations. The same then applies to indirect objects: in HUDDv2, 60 iobj follow the verb they depend on and 49 precede it, for a VI to IV ratio of 55% to 45%. In Maltese, as evident from the data in Table 7.16, the same ratio is 85.7% to 14.3%.

7.4.3 Conclusion

One might argue that this little comparison does not prove very much: for one, both treebanks are relatively small and thus hardly representative of the language as a whole, especially seeing as the Hungarian UD v2 treebank only includes journalistic

texts (Nivre, Agić et al. 2017). Additionally, Fabri (2010: 793) may very well be correct in arguing that spoken Maltese is different from written Maltese when it comes to constituent order and so a treebank consisting of spoken materials only might offer a different picture.

As a rebuttal for the second objection, I offer this back-of-the-envelope calculation: MUDTv1 contains 1911 clauses featuring a dobj or a nmod:obj,[12] of which 90 are OV, for a rate of 4.7%; the same rate is 53.1% for HUUDv2. If one were to add 100[13] OV clauses to every UD clause type in MUDTv1 thus increasing the total count of OV clauses to 790, the overall OV share in MUDTv1 would climb to only 30% and, as evident from Figure 7.23, it would still barely approach the level of OV in HUUDv2.

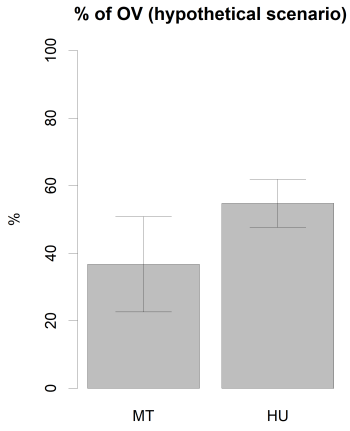


Fig. 7.23: MUDTv1 vs HUUDv2: VO vs OV order with 100 additional OV clauses in MUDTv1 per clause type

It would therefore seem more likely that MUDTv1 represents this particular aspect of Maltese as a whole rather faithfully (in other words, spoken Maltese may very well be different from written Maltese, but it surely isn't that different) and that the differences between the two treebanks really are significant. And, to answer the first objection, the composition of HUUDv2 only underscores this: journalistic texts are typically written in a dry and formal style driven by desire for clarity and brevity and produced under time crunch, which encourages the use of canned constructions ("journalese", Suter 1993:

12 This excludes the four dobj in csubj, as well as the three cases of an X_FOR governing a dobj (see Table 7.4).

13 I.e. as many OV clauses as there are in MUDTv1 now, plus 10 more just to get a nice round number.

63:-68). The fact that even when compared to a relatively balanced MUDTv1, HUUDv2 is so different when it comes to the distribution of VO and OV configurations, then cannot be explained away by sampling issues. This is doubly true in light of the fact that (as evident from Table 7.27) if one were to compare journalistic texts only, the difference would be even more pronounced: in those types of texts in MUDTv1, the share of the OV configuration (2.8%) is even lower than the average in MUDTv1 (4.7%).

Consequently, there are two conclusions to be drawn here: first, Maltese (at least as represented in MUDTv1) really is fundamentally different from Hungarian (as represented in HUUDv2) when it comes to the distribution of constituent order configurations and ipso facto, the two languages cannot belong to the same class defined by a shared property related to constituent order. If one chooses to describe Hungarian as a discourse-configurational language based on the description of its constituent order, it does not seem appropriate to do the same for Maltese. By extension, neither does applying the label "topic-prominent".

The second conclusion to be drawn from the calculations above is essentially the same as the first one, except broader and methodological rather than descriptive: Borg and Fabri use the "discourse-configurational" label as a typological one which is itself somewhat problematic. The real problem, however, is that they do so without considering the entire theory it is based on.[14] As a part of a generative framework, discourse-configurationality is inexorably tied to its fundamental theory of sentence production and its complex conceptual apparatus including base generation, movements and functional projections (cf. Kiss 1995b: 9-10). And even if they were to argue that they only borrow the name and the descriptive information structure concepts behind it (as opposed to the theory of sentence generation), Borg and Fabri fail to consider one crucial property of discourse-configurational languages as defined by Kiss (1995b), the empirical distinction between categorical and thetic statements. In Kiss's wider definition, "[a] language is identified as topic-prominent, more precisely, as a discourse configurational language with property A, if it realizes categorical and thetic judgements in different syntactic structures" (Kiss 1995b: 7-8, see also Chapter 2). Their work does not take this into account and this further invalidates their description of Maltese as a discourse-configurational or a topic-prominent language: such a label, after all, only makes sense within the context of the theory.

Ironically, I've shown here that Maltese actually does employ a different syntactic structure for at least one type of thetic judgments, existential clauses, so taking this into account would support Fabri and Borg's description of Maltese as discourse-configurational as defined in the theory. This argument could be used to make a renewed case for this classification. One could, for example, extend the comparison provided here to other languages and offer data such as those in Figure 7.25 and 7.24.

14 This is not the case with Fabri (1993: 140) who describes Maltese a configurational language, citing the exact definition established in generative literature.

These plots were made using data from MUDTv1 and three UD v2.1 treebanks (Nivre, Agić et al. 2017): two for languages considered discourse-configurational, Hungarian and Modern Greek (Kiss 1995b: 6),[15] and one for a language with a rigid SVO order (Dryer 2013c) and the very opposite of a discourse-configurational language (Kiss 1995: 5, 8), English.

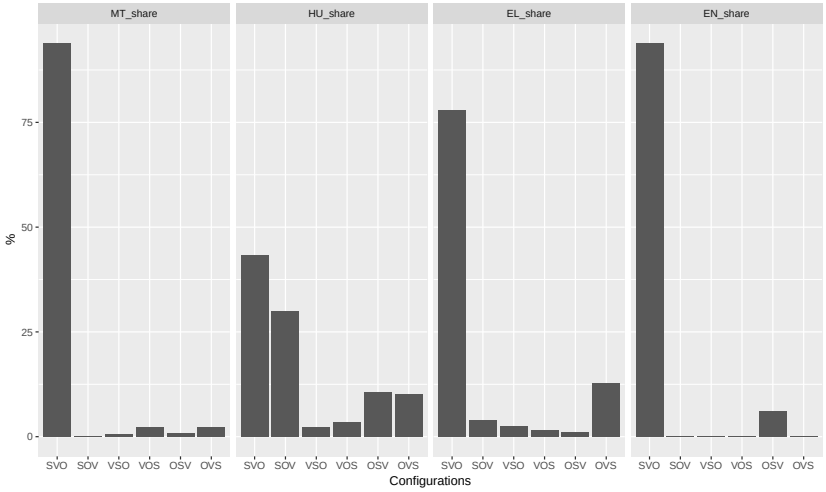


Fig. 7.24: Greenbergian classification of Maltese, Hungarian, Greek and English

15 See also WALS characterization of both Greek and Hungarian as languages with no dominant constituent order (Dryer 2013c).

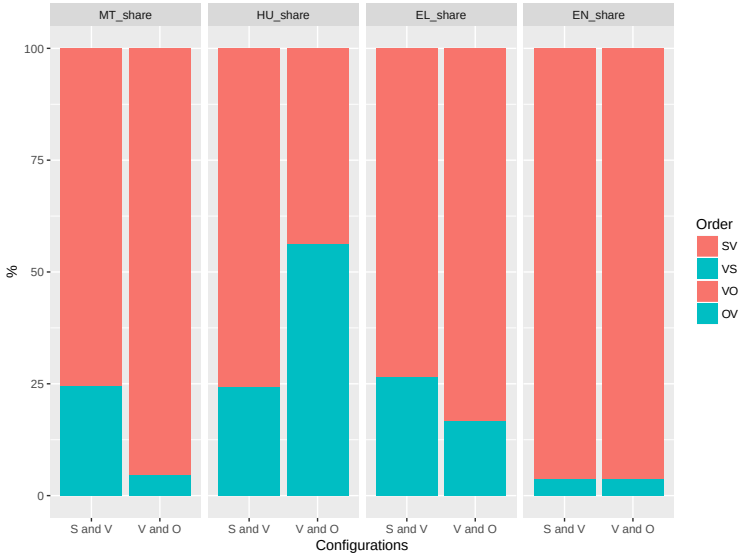


Fig. 7.25: Dryerian classification of Maltese, Hungarian, Greek and English

Upon reviewing this data, one could observe that Hungarian and English behave quite differently, as expected from their respective typological classifications. One could also note that Greek, another discourse-configurational language, is also quite different from English, yet at least when it comes to the order of object and predicate, it also behaves differently from Hungarian. And so one could argue that discourse-configurationality is a scale, with Hungarian on one end and English on the other and Greek and Maltese somewhere in between; though obviously Maltese is in one aspect (the order of V and O) closer to English[16] than Greek. Whether such a conclusion is consistent with the definition of discourse-configurationality as provided by Kiss (1995a), will be left to that hypothetical observer. I, for one, will remain silent on the question of the validity of the theory, but continue to question the classification Maltese as discourse-configurational on the basis of the data above.

16 Assuming, of course, the OV clauses in English are not all interrogatives, which may very well be the case.

7.5 Summary

7.5.1 Introduction

In this section, answers are provided to the Research Questions in the order that they were asked.

7.5.2 Answer to Research Question 1: What is the dominant constituent order in Maltese?

The dominant constituent order in Maltese is SV (except for the variation described in the next section) and VO. The dominant constituent order in Maltese can also be described as SVO.

7.5.3 Answer to Research Question 2: What is the variation in dominant constituent order in Maltese?

In Maltese, the dominant order of subject and predicate is VS rather than SV in all existential clauses and in verbal parataxis clauses; no dominant order could be established for verbal acl clauses (both active and passive), passive advcl clauses, passive ccomp clauses and copular xcomp clauses. Table 7.28 contains a summary with the variation highlighted in bold.

UD clause type	Verbal (active)	Verbal (passive)	Copular	Existential
main	SV	SV	SV	**VS**
acl	**no dominant order**	**no dominant order**	SV	**VS**
advcl	SV	**no dominant order**	SV	**VS**
xcomp	SV	n/a	**no dominant order**	**VS**
ccomp	SV	**no dominant order**	SV	**VS**
parataxis	**VS**	SV	SV	**VS**
conj	SV	SV	SV	**VS**

Tab. 7.28: MUDTv1: Variation in dominant constituent order

In contrast, the dominant order of predicate and object is VO in all clause types.

7.5.4 Answer to Research Question 3: What are the deviations from the dominant constituent order in Maltese?

The deviations from the dominant constituent order are recorded in plots and tables in section 7.3.

Of particular interest is the share of OV clauses (see Table 7.13 and the analysis of the deviation in Table 7.14). As noted in section 7.3.2.3, these fall into a number of distinct groups, one of which involves what previous descriptions of constituent order of Maltese refer to as topicalization of objects (Borg and Azzopardi-Alexander 1997, 2009). These constructions have been described as "a wide spread characteristic of Maltese" (Borg and Azzopardi-Alexander 1997: 126) and yet in MUDTv1, only 48 clauses, i.e. 2.5% of all direct objects, fall into that group (interestingly, the ratio for topicalized indirect objects is nearly identical at 2.6%). One might once again justifiably question the balance and representativeness of MUDTv1 or invoke the special nature of spoken language in explaining the discrepancy between the MUDTv1 data and the descriptions such as the one above, but until more data is available, the conclusion to be drawn here is that in Maltese, topicalization of objects is not as wide-spread as generally thought.

7.5.5 Answer to Research Question 4: What are the factors that cause variation in dominant constituent order?

Existential clauses are the only type of clause that consistently exhibits dominant order opposite to that of the other clauses and Maltese as a whole. The dominant predicate-first order reflects what has been observed regarding existential clauses cross-linguistically (cf. Givón 2001b: 257, McNally 2011: 1833). Existing hypotheses explain this in terms of information structure or in terms of diachronic development (Givón 2001b: 259); whether that holds for Maltese remains to be seen.

For those clause types where the dominant order could not be determined (Table 7.28), preliminary research indicates that heaviness of the subject and length of the clause play a role with the former being associated with VS and the latter with SV. Whether this is in fact true and if, what other factors there are and how they interact with heaviness and length, is a question that remains to be answered.

7.5.6 Final considerations

In this chapter, I confronted the description of Maltese as a discourse-configurational language with data from MUDTv1 and its equivalent from the Hungarian UD v2.1 treebank. Based on this, I concluded that Hungarian is quite different from Maltese when it comes to constituent order variation and so if one were to apply the label "discourse-

configurational" (or indeed any other) to Hungarian based on the properties of its constituent order, one should not do so for Maltese.

Whether I am justified in arguing that and, more generally, whether empirical data like this can invalidate a description grounded in a particular framework (theory of sentence production) is a question that goes to the heart of what linguistics is and what is its purpose, one that I addressed in Chapter 1 and one that will ultimately have to be answered by minds superior to mine. As my own contribution to answering it and at the same time a conclusion to this aside, I will recall another passage in Chapter 1 in which I discussed the term "description". This I employed to define my approach to studying language, noting that it is equivalent to Martin Haspelmath's (2004) "phenomenological description", but with the proviso that Haspelmath argues that phenomenological description entails accurate prediction of speaker behavior (Haspelmath 2009: 344). I found that particular part of his definition troubling, questioning the utility of prediction vis-à-vis such a complex and chaotic system as a human being. I still hold that view, but (being now much wiser than I was when I wrote chapter 1) perhaps less rigidly so, seeing the validity of some of the arguments brought forth by Haspelmath 2009 (and Gries 2001 and Köhler 2012: 14-15) in this regard. And so when in section 7.5.1 I spoke of expecting the share of SV to VS and VO to OV to be roughly the same in a discourse-configurational language, I was actually making a prediction about what data gathered for a discourse-configurational language would look like or a prediction about what types of sentences speakers of a discourse-configurational language would produce. As it turned out, the prediction was correct for sentences produced by speakers of Hungarian (as represented in HUUDv2), but incorrect for sentences produced by speakers of Maltese (as represented in MUDTv1), at least as far as the order of V and O is concerned.

Haspelmath's discussion of phenomenological description and the role of prediction in it concludes with the following: "Thus linguists must by and large be content with descriptions that accurately predict the behavior of speakers in natural corpora and experimental contexts" (Haspelmath 2009: 344). Accepting this as a goal of linguistics as a science has a profound effect on the philosophy and methodology of linguistics: to make predictions, one needs a conceptual apparatus capable of making predictions. In the context of constituent order, Dryer's SV/VS and VO/OV typology is such a conceptual apparatus: one can use it (with requisite degree of caution) to make predictions as to what particular configuration will speakers of a particular language use in what contexts; the same applies (with even more caution) to Greenbergian six-way typology. Whether the theory of discourse configurationality is such a conceptual apparatus is debatable: I have used it as one, but only by naively inferring from the theory, as the theory itself is concerned with the development of the framework rather than description (which is another reason why using "discourse-configurational" as a typological label is inadvisable). What is certain, however, is the fact that descriptions of constituent order like "free" or "pragmatically determined" are wholly incapable of predicting any-

thing. At the very best, they indicate potentiality, but nothing more; as such, they are nigh useless for description, even if used in contrast.

And this brings me to my last point and two more concepts harking back to Chapter 1: the first is the distinction between descriptive and theoretical linguistics (discussed in section 1.2), where the former has description (and, yes, prediction) as the goal, while the latter seeks to create an accurate representation of the speakers' mental grammars. It is precisely this distinction that Dryer refers to when he speaks of the role of frequency in the study of constituent order: on one hand, he insists that "the greater frequency of [configuration] and [other configuration] orders is ... not a fact about the grammar of [language]" (Dryer 1997: 73) while making it clear that by "grammar" he means "mental representation of language" ("speakers store grammatical knowledge independent of frequency", Dryer 2013b: 292). On the other, Dryer regularly cites frequency as one of the major factors "in explaining why languages — and grammars — are the way they are" (Dryer 2013b: 292). That last citation comes from a paper in which he argues against a generativist defense of Greenbergian six-way typology (Newmeyer 2005) offered in the context of defending the distinction between grammar and usage.[17] Here Dryer backs down or wants to have it both ways, and so he does not follow his line of thinking through by arguing that frequency is a part of usage and thus usage determines grammar (if only diachronically). I feel inclined to do so, but — as discussed in Chapter 1 — being an empirical nominalist, I don't have to. And so will satisfy myself by pointing out the following: that speakers of Maltese can (as shown above and in Fabri and Borg 2002) and do produce OV&VS sentences is a fact about Maltese. That in ~75% of cases these speakers produce clauses with SV order and in ~95% of cases they produce clauses with VO order (as shown above) is also a fact about Maltese. Both these facts need to be reconciled in a way which not only accurately describes data, but allows for making correct predictions, regardless of whether or not these two facts belong to different linguistic domains.

17 Newmeyer (2005: 161), where, incidentally, Newmeyer also invokes prediction by arguing that "No generative grammarian ever claimed that sentences generated by the grammar should be expected to reveal directly what language users are likely to say" (Newmeyer 2005: 152).

8 Summary

8.1 The road up here

The primary contribution of this work to the study of Maltese, the study of syntax, and linguistics in general, is two-fold: first, I have described and made available to the public two sets of language resources for Maltese, a general corpus (*BCv3*, Chapter 5) and a syntactically annotated corpus (treebank) of written Maltese (MUDTv1, Chapter 6). And while the former has actually been around for a few years (and has been used to provide insight into several aspects of Maltese syntax) and thus it is the detailed description of its composition and annotation that is new, the Maltese Universal Dependencies Treebank v1 (based on the Universal Dependencies standard) is the first such resource for Maltese. The compilation of such a resource is a complex undertaking in and of itself: the selection of texts and text types, reflecting the narrow options afforded to those working on Maltese, nevertheless aimed at compensating for the opportunistic and imbalanced nature of *BCv3* (and MLRSv3, the other general corpus of Maltese).

The actual annotation of the treebank consists of nothing short of a sketch of Maltese syntax based on the principles of dependency linguistics. As such, the annotation of MUDTv1 required refining the existing analyses of several phenomena (e.g. copular clauses, auxiliary verbs, the phenomenon of verbal chains and verbal complementization) while providing a new analysis to phenomena not yet or sufficiently accounted for, chief among them the classification of clauses by root and the resulting clause structure, the identification of several types of existential clauses and most importantly, a preliminary analysis of verbal valency in Maltese. Using the Valency Lexicon of Czech Verbs (VALLEX) as a model, a preliminary classification of Maltese verbal dependents is laid out which, inter alia, identifies a type of obligatory dependent (and thus a type of trivalent verbs) hitherto not discussed in literature on Maltese grammar. Although an in-depth discussion of the valency of Maltese verbs is beyond the scope of this work, a groundwork is nevertheless laid for the same, while also providing a quick and consistent way of identifying and classifying verbal dependents for the purposes of syntactic annotation.

The second major contribution of this work is that stated in its title, i.e. the investigation of constituent order in Maltese using quantitative methods on MUDTv1 data (Chapter 7). The results of this investigation confirm the general typological classification of Maltese as SVO (where the share of SVO clauses is 93.9%) and SV/VO (~75% for SV clauses, ~95% for VO clauses). Some variation has been found, such as the lack of any dominant order in some type of UD clauses (primarily acl), as well as VS as the dominant constituent order in existential clauses which is consistent with cross-linguistic observations regarding these types of clauses. For those clauses where no dominant constituent order could be found, statistical modeling was employed to account for it and while the results of said analysis are preliminary at best, they point to heaviness

and clause length playing a role, the former associated with VS order, the latter with SV.

Additionally, data from other UD treebanks, primarily Hungarian, was used to test the hypothesis (as offered by description provided by some previous works) that Maltese is a discourse-configurational language by comparing it to Hungarian, the very definition of a discourse-configurational, as well as to other languages. This comparison showed that at least when it comes to the ordering of the object and the predicate in verbal clauses, Maltese (as represented in MUDTv1) is significantly different from that of Hungarian and both should thus not be described as belonging to the same class of languages defined by a shared property involving constituent order. Expanding on that observation, this work confirms not only the utility of Dryer's SV/VS and VO/OV typological classification for the description of constituent order, but also the inutility of classifications such as "free word order" or "discourse-configurational". This evaluation is based, inter alia, on the fact that while previous literature on Maltese describes OV order as a wide-spread and conspicuous feature of Maltese, such constructions are in fact quite rare, occurring only in 2.5% of all clauses featuring a direct object.

8.2 The road ahead

It should go without saying that the conclusions presented in this work are far from the last word on the subject. In fact, the exact opposite is true: the data and the insights drawn from them offered here should be viewed as nothing but a first step towards the detailed description of constituent order in Maltese and its variation (and deviation). As evident from the many instances of "beyond the scope of this work" and its synonyms dispersed throughout the previous 250+ pages, many of the issues involved are complex and requiring extensive treatment, preferably on more and more varied data. Chief among them is the issue of actual spoken Maltese and to what extent it is (if at all) different from the written language represented in MUDTv1. Answering this question would, naturally, involve building and annotating a treebank of spoken Maltese and while some steps towards that goal have been taken (Paggio and Vella 2014), such a treebank is still in our future; so is the expansion of MUDTv1 with more and more diverse texts which would make a follow-up confirmation study possible.

Let us therefore focus on the many shortcomings of this work and how they can be addressed using the data already available. Firstly, with the exception of the analysis of the deviant OV order in Chapter 7, section 7.3.2.3, the analysis here largely remains silent on constituent order in sentence types classified by modality (imperative, exhortative, interrogative etc.). This is of course a glaring omission as in Maltese, some types of questions have been shown to display deviations from the dominant order, especially as far as objects are concerned (Borg and Azzopardi-Alexander 1997: 20). Secondly, while Chapter 7, section 7.3.2.3 also attempts to account for the OV deviation in terms of information structure, this analysis should be first put on a more solid theoret-

ical footing by providing clear and actionable definitions of the information structure concepts in question and then expanded to the VS (or SV, in case of existential clauses) deviation. And of course with further effort at annotating semantic properties of core dependents (e.g. animacy), analyses like the one conducted for acl clauses in 7.3.2.2.4 can be extended to all types of clauses to determine the full set of factors influencing constituent order variation and deviation.

Going beyond the analysis of constituent order, the very next project MUDTv1 should be used for is the analysis of word order and its relationship to constituent order, clause structure and complex sentence structure, including, but not limited to, a full classification of branchedness in Maltese. The question of the order of elements within a noun phrase is a particularly fascinating one, as it touches upon the mixed nature of Maltese morphology and syntax. Such an analysis can then be immediately extended to that of valency of nouns and adjectives while reviewing and expanding the work begun here on verbal valency and providing a more generally grounded description of non-canonical objects, including their passive diathesis. This, naturally, ties to the further development of the treebank where a number of areas not involving constituent or word order need to be revisited; these include paratactic clauses and their further subdivision, comparative constructions, numerals, compounds (especially Light Verb Constructions), xcomp clauses featuring pseudoverbs, the concept of "auxiliary verb" in Maltese, as well as further refinement and classification of the generic nmod relation used for nominal dependents of noun phrases.

And finally, this thesis has pointed out a number of issues of general descriptive import. Chapter 6 highlighted several lacunae in the description of Maltese, such as the copular and existential clauses, which both (but especially the latter) lack a comprehensive synchronic and diachronic description, as well as an analysis of their areal aspects (Arabic varieties for type (ii) and especially type (iii) copular clauses;[1] both Arabic and Romance neighbors and ancestors of Maltese for existentials). The same is of course true of non-copular verbless clauses and the synchrony and diachrony of the expletive, as well as of non-expletive subjectless clauses. Chapters 6 and 7 also demonstrated that verbal chains and verbal complementation in Maltese still lack a satisfactory description; the same applies to various dislocation phenomena, coordination and ellipsis.

The data provided and described in this work can be used to accomplish these goals and even more. Onward and upward!

1 Indeed one such study on the latter clause type (Camilleri and Sadler 2018) was made public just as finishing touches were being put to this work.

Bibliography

Ágel, Vilmos et al., eds. (2003). *Dependenz und Valenz / Dependency and Valency. Ein internationales Handbuch der zeitgenössischen Forschung / An International Handbook of Contemporary Research. 1. Halbband / Volume 1*. Berlin / New York: Walter De Gruyter.

Andor, Daniel et al. (2016). "Globally Normalized Transition-Based Neural Networks". In: *CoRR* abs/1603.06042.

Antomonov, Filip (2015). "UD: a parsing comparison".

Aquilina, Joseph (1959). *The Structure of Maltese: A Study in Mixed Grammar and Vocabulary*. Malta: The Royal University of Malta.

— (1976). *Maltese Linguistic Surveys*. Malta: The University of Malta.

Arnold, Jennifer E. et al. (2000). "Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering". In: *Language* 76.1, pp. 28–55.

Baayen, R. H. (2008). *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge University Press.

Bader, Markus and Jana Häussler (2010). "Word order in German: A corpus study". In: *Lingua* 120.3, pp. 717–762.

Bailey, Laura R. and Michelle Sheehan, eds. (2017). *Order and structure in syntax I: Word order and syntactic structure*. Vol. I. Berlin: Language Science Press.

Bakker, Dik (1998). "Flexibility and consistency in word order patterns in the languages of Europe". In: *Eurotyp: Typology of Languages in Europe, Volume 1: Constituent Order in the Languages of Europe*. Ed. by Anna Siewierska, pp. 383–419.

Bates, Douglas et al. (2015). "Fitting Linear Mixed-Effects Models Using lme4". In: *Journal of Statistical Software* 67.1, pp. 1–48.

Bejček, Eduard et al. (2013). *Prague Dependency Treebank 3.0*. Data/software. Prague: Univerzita Karlova v Praze, MFF, ÚFAL.

Benincà, Paola (2006). "A Detailed Map of the Left Periphery of Medieval Romance". In: *Crosslinguistic Research in Syntax and Semantics: Negation, Tense, and Clausal Architecture*. Ed. by Raffaella Zanuttini et al. Washington, D.C.: Georgetown University Press, pp. 53–86.

Benincà, Paola and Nicola Munaro (2010a). "Introduction". In: *Mapping the Left Periphery: The Cartography of Syntactic Structures, Volume 5*. Ed. by Paola Benincà and Nicola Munaro. Oxford: Oxford University Press, pp. 3–15.

— eds. (2010b). *Mapping the Left Periphery: The Cartography of Syntactic Structures, Volume 5*. Oxford: Oxford University Press.

Benincà, Paola and Cecilia Poletto (2004). "Topic, Focus, and V2: Defining the CP Sublayers". In: *The Structure of CP and IP: The Cartography of Syntactic Structures Volume 2*. Oxford: Oxford University Press, pp. 52–75.

Benkato, Adam and Christophe Pereira (2015). "bda et g^{c}əd en arabe de Tripoli et en arabe de Benghazi (Libye)". A paper read at the 11th conference of AIDA, Bucharest, May 25 – May 28 2015.

Bentley, Delia (2015a). "Existentials and Locatives in Romance Dialects of Italy. Introduction". In: *Existentials and Locatives in Romance Dialects of Italy*. Ed. by Delia Bentley, Francesco Maria Ciconte, and Silvio Cruschina. Oxford: Oxford University Press, pp. 1–42.

— (2015b). "Predication and argument realization". In: *Existentials and Locatives in Romance Dialects of Italy*. Ed. by Delia Bentley, Francesco Maria Ciconte, and Silvio Cruschina. Oxford: Oxford University Press, pp. 99–160.

Bentley, Delia, Francesco Maria Ciconte, and Silvio Cruschina, eds. (2015). *Existentials and Locatives in Romance Dialects of Italy*. Oxford: Oxford University Press.

Berlinches, Carmen (2016). *El dialecto árabe de Damasco (Siria): estudio gramatical y textos*. Estudios de Dialectología Árabe 11. Zaragoza: Prensas de la Universidad de Zaragoza.

Bielický, Viktor (2015). "Valenční slovník arabských sloves. Valency Dictionary of Arabic Verbs". PhD thesis. Charles University in Prague.

Boeckx, Cedric (2006). *Linguistic Minimalism: Origins, Concepts, Methods, and Aims*. Oxford: Oxford University Press.

Borg, Albert (1987/88). "To be or not to be: A copula in Maltese". In: *Journal of Maltese Studies* 17/18, pp. 54–71.

Borg, Albert and Marie Azzopardi-Alexander (1997). *Maltese*. London: Routledge.

— (2009). "Topicalisation in Maltese". In: *Introducing Maltese Linguistics*. Ed. by Bernard Comrie et al. Amsterdam: John Benjamins, pp. 71–81.

Borg, Albert and Ray Fabri (2016). "Semantic functions of complementizers in Maltese". In: *Complementizer Semantics in European Languages*. Ed. by Kasper Boye and Petar Kehayov. Berlin/Boston: De Gruyter Mouton, pp. 413–447.

Borg, Albert and Michael Spagnol (2015). "Nominal sentences in Maltese: a corpus-based study". A paper read at the 5th International Conference on Maltese Linguistics, June 24-26 2015, L'Università di Torino.

Borg, Alexander (1985). *Cypriot Arabic: a historical and comparative investigation into the phonology and morphology of the Arabic vernacular spoken by the Maronites of Kormakiti village in the Kyrenia district of North-West Cyprus*. Stuttgart: Deutsche Morgenländische Gesellschaft.

Borg, Claudia (2016). "Morphology in the Maltese Language: A Computational Perspective". PhD thesis. University of Malta.

Bovingdon, R. and A. Dalli (2006). "Statistical analysis of the source origin of Maltese". In: *Corpus linguistics around the world*. Ed. by A. Wilson, D. Archer, and P. Rayson. Amsterdam: Rodopi, pp. 63–76.

Boye, Kasper and Petar Kehayov, eds. (2016). *Complementizer Semantics in European Languages*. Berlin/Boston: De Gruyter Mouton.

Broekhuis, Hans (2008). *Derivations and Evaluations: Object Shift in the Germanic Languages*. Studies in generative grammar 97. Berlin / New York: Mouton de Gruyter.

Camilleri, John J. (2013). "A Computational Grammar and Lexicon for Maltese". MA thesis. Gothenburg, Sweden: Chalmers University of Technology.

Camilleri, Maris (2016). "Temporal and Aspectual auxiliaries in Maltese". PhD thesis. University of Essex.

— (2018). "Clausal possession in Maltese". In: *Studies in Language*.

Camilleri, Maris and Louisa Sadler (2016). "Relativisation in Maltese". In: *Transactions of the Philological Society* 114.1, pp. 117–145.

— (2018). "The gramaticalisation of a stage-level copula in vernacular Arabic". Paper presented at the 32nd Annual Symposium on Arabic Linguistics, Arizona State University, February 23-25, 2018.

Canty, Angelo and B. D. Ripley (2017). *boot: Bootstrap R (S-Plus) Functions*. R package version 1.3-20.

Carnie, Andrew (2013). *Syntax: A Generative Introduction*. Chichester: Wiley-Blackwell.

Caruana, Sandro, Ray Fabri, and Thomas Stolz, eds. (2011). *Variation and Change. The Dynamics of Maltese in Space, Time and Society*. Studia Typologica 9. Berlin: Akademie Verlag.

Čéplö, Slavomír (2014). "An overview of object reduplication in Maltese". In: *Perspectives on Maltese Linguistics*. Ed. by Albert Borg, Sandro Caruana, and Alexandra Vella. Berlin: De Gruyter, pp. 201–222.

— (2017). "Focus particles in Maltese: A corpus survey". In: *Advances in Maltese Linguistics*. Ed. by Benjamin Saade and Mauro Tosco. Berlin: De Gruyter, pp. 87–120.

Čéplö, Slavomír and Lonneke van der Plas (2017). "Light verb constructions in Maltese: Evidence from corpora". A paper read at the Sixth International Conference on Maltese Linguistics, June 8th – June 9th, Comenius University in Bratislava.

Choe, Hyoon Sook (1995). "Focus and Topic Movement in Korean and Licensing". In: *Discourse Configurational Languages*. Ed. by Katalin É. Kiss. Oxford: Oxford University Press, pp. 269–334.

Chomsky, Noam (1964). *Current Issues in Linguistic Theory*. Janua Linguarum, Series Minor 38. The Hague / Paris: Mouton.

— (1965). *Aspects of the Theory of Syntax*. Cambridge, Mass: MIT Press.

— (1981). *Lectures on Government and Binding: The Pisa Lectures*. Dordrecht - Holland: Foris Publications.

— (1985). *Barriers*. Linguistic Inquiry Monographs 13. Cambridge, Massachusetts: The MIT Press.

— (1995). *The Minimalist Program*. Current Studies in Linguistics. Cambridge, Massachusetts / London, England: The MIT Press.

Chomsky, Noam and Mitsou Ronat (1979). *Language and Responsibility*. Pantheon Books.

Cinque, Guglielmo, ed. (2002a). *Functional Structure in DP and IP: The Cartography of Syntactic Structures, Volume 1*. Oxford: Oxford University Press.

— (2002b). "Mapping Functional Structure: A Project". In: *Functional Structure in DP and IP: The Cartography of Syntactic Structures, Volume 1*. Ed. by Guglielmo Cinque. Oxford: Oxford University Press, pp. 3–11.

Cinque, Guglielmo and Luigi Rizzi (2010). "The Cartography of Syntactic Structures". In: *The Oxford Handbook of Linguistic Analysis*. Ed. by Bernd Heine and Heiko Narrog. Oxford: Oxford University Press, pp. 51–65.

Cognola, Federica (2013). *Syntactic Variation and Verb Second: A German dialect in Northern Italy*. Amsterdam/Philadelphia: John Benjamins.

Comrie, Bernard (1989). *Language Universals and Linguistic Typology (2nd edition)*. Chicago: The University of Chicago Press.

— (2009). "Maltese and the World Atlas of Language Structures". In: *Introducing Maltese Linguistics*. Ed. by Bernard Comrie et al. Amsterdam: John Benjamins, pp. 3–14.

Comrie, Bernard and Norval Smith (1977). "Lingua descriptive studies: Questionnaire". In: *Lingua* 42.1, pp. 1–71.

Corver, Norbert and Henk van Riemsdijk (1994a). "Introduction: approaches to and properties of scrambling". In: *Studies on Scrambling: Movement and Non-Movement Approaches to Free Word-Order Phenomen*. Ed. by Norbert Corver and Henk van Riemsdijk. Berlin / New York: Mouton de Gruyter, pp. 1–13.

— eds. (1994b). *Studies on Scrambling: Movement and Non-Movement Approaches to Free Word-Order Phenomen*. Studies in Generative Grammar. Berlin / New York: Mouton de Gruyter.

Culicover, Peter W. and Ray Jackendoff (2005). *Simpler Syntax*. Oxford University Press.

Cvrček, Václav et al. (2015). *Mluvnice současné češtiny I. Jak se píše a jak se mluví*. Praha: Karolinum.

Dalrymple, Mary (2001). *Lexical Functional Grammar*. Syntax and Semantics Series, Volume 34. New York: Academic Press.

Danckaert, Lieven (2017). *The Development of Latin Clause Structure: A Study of the Extended Verb Phrase*. Oxford: Oxford University Press.

Daneš, František (1957). *Intonace a věta ve spisovné češtině*. Praha: ČSAV.

— (1959). "K otázce pořádku slov v slovanských jazycích". In: *Slovo a slovesnost* 20.1, pp. 1–10.

Davison, A. C. and D. V. Hinkley (1997). *Bootstrap Methods and Their Applications*. ISBN 0-521-57391-2. Cambridge: Cambridge University Press.

de Cat, Cécile (2010). *French dislocation. Interpretation, syntax, acquisition*. Oxford: Oxford University Press.

de Marneffe, Marie-Catherine, Timothy Dozat, et al. (2014). "Universal Stanford Dependencies: a Cross-Linguistic Typology". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Ed. by Nicoletta Calzolari (Conference Chair) et al. Reykjavik, Iceland: European Language Resources Association (ELRA), pp. 26–31.

de Marneffe, Marie-Catherine, Bill MacCartney, and Christopher D. Manning (2006). "Generating Typed Dependency Parses from Phrase Structure Parses". In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-2006)*. ACL Anthology Identifier: L06-1260. Genoa, Italy: European Language Resources Association (ELRA).

de Marneffe, Marie-Catherine and Christopher D. Manning (2008). "The Stanford Typed Dependencies Representation". In: *Coling 2008: Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation*. CrossParser '08. Manchester, United Kingdom: Association for Computational Linguistics, pp. 1–8.

de Soldanis, Giovanni Pietro Francesco Agius (1750). *Della lingua punica presentemente usata da maltesi*. Rome.

Derbyshire, Desmond C. (1977). "Word Order Universals and the Existence of OVS Languages". In: *Linguistic Inquiry* 8.3, pp. 590–599.

Dixon, R. M. W. (2009). *Basic Linguistic Theory Volume 1: Methodology*. Oxford: Oxford University Press.

Dixon, R. M. W. and Alexandra Y. Aikhenvald, eds. (2003). *Word: A Cross-linguistic Typology*. Cambridge: Cambridge University Press.

Dryer, Matthew S. (1989a). "Discourse-Governed Word Order and Word Order Typology". In: *Belgian Journal of Linguistics* 4, pp. 69–90.

— (1989b). "Large linguistic areas and language sampling". In: *Studies in Language* 13, pp. 257–292.

— (1992). "The Greenbergian Word Order Correlations". In: *Language* 68.1 (1), pp. 81–138.

— (1997). "On the 6-way Word Order Typology". In: *Studies in Language*, pp. 69–103.

— (2007). "Word order". In: *Language Typology and Syntactic Description. Second edition. Volume I: Clause Structure*. Ed. by Timothy Shopen. Vol. I. Cambridge: Cambridge University Press, pp. 61–131.

— (2013a). "Determining Dominant Word Order". In: *The World Atlas of Language Structures Online*. Ed. by Matthew S. Dryer and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology, n/a.

— (2013b). "On the Six-Way Word Order Typology, Again". In: *Studies in Language* 37.2, pp. 267–301.

— (2013c). "Order of Subject, Object and Verb". In: *The World Atlas of Language Structures Online*. Ed. by Matthew S. Dryer and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Dürscheid, Christa (2007). *Syntax. Grundlagen und Theorien*. 4., überarbeitete und ergänzte Auflage. Göttingen: Vandenhoeck & Ruprecht.

Dušková, Libuše (2015). *From Syntax to Text: the Janus face of Functional Sentence Perspective*. Praha: Karolinum.

Fabri, Ray (1993). *Kongruenz und die Grammatik des Maltesischen*. Tübingen: Max Niemeyer Verlag.

— (2005). "Clitics and non-definite NPs in Maltese". A paper read at the The Third International Conference on Maltese Linguistics, April 8-10 2011, University of Malta.

— (2010). "Maltese". eng. In: *Revue belge de philologie et d'histoire* 88.3, pp. 791–816.

Fabri, Ray and Albert Borg (2002). "Topic, Focus and Word Order in Maltese". In: *Aspects of the Dialects of Arabic Today*. Ed. by A. Youssi and Alii. AIDA. Rabat: AMAPATRIL, pp. 354–363.

— (2017). "Modifiers and complements within the Maltese verb sequence". In: *Advances in Maltese Linguistics*. Ed. by Benjamin Saade and Mauro Tosco. Berlin: De Gruyter, pp. 67–86.

Fenech, Edward (1978). *Contemporary journalistic Maltese*. Leiden: E. J. Brill.

Féry, Caroline and Shinichiro Ishihara (2016a). "Introduction". In: *The Oxford Handbook of Information Structure*. Ed. by Caroline Féry and Shinichiro Ishihara. Oxford: Oxford University Press, pp. 3–15.

— eds. (2016b). *The Oxford Handbook of Information Structure*. Oxford: Oxford University Press.

Féry, Caroline and Manfred Krifka (2008). "Information structure: Notional distinctions, ways of expression". In: *Unity and Diversity of Languages*. Ed. by Piet van Sterkenburg. Amsterdam / Philadelphia: John Benjamins, pp. 123–135.

Firbas, Jan (1964). "On Defining the Theme in Functional Sentence Analysis". In: *Travaux Linguistiques de Prague* 1, pp. 267–280.

— (1992). *Functional sentence perspective in written and spoken communication*. Cambridge: Cambridge University Press.

Freidin, Robert and Howard Lasnik (2011). "Some Roots of Minimalism in Generative Grammar". In: *The Oxford Handbook of Linguistic Minimalism*. Ed. by Cedric Boeckx. Oxford University Press, pp. 1–26.

Gallego, Ángel J. (2013). "Object shift in Romance". In: *Natural Language & Linguistic Theory* 31.2, pp. 409–451.

Gatt, Albert and Slavomír Čéplö (2013). "Digital corpora and other electronic resources for Maltese". Talk at the Corpus Lingustics 2013 conference at Lancaster University.

Gerdes, Kim and Sylvain Kahane (2011). "Defining dependencies (and constituents)". In: *depling 2011 proceedings*. Ed. by Eva Hajičová Kim Gerdes and Leo Werner. Barcelona: depling.org, pp. 17–27.

Giménez, Jesús and Lluís Márquez (2004). "SVMTool: A general POS tagger generator based on Support Vector Machines". In: *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pp. 43–46.

Givón, Talmy (2001a). *Syntax. An Introduction*. Vol. I. Amsterdam/Philadelphia: John Benjamins.

— (2001b). *Syntax. An Introduction*. Vol. II. Amsterdam/Philadelphia: John Benjamins.

Graffi, Giorgio (2001). *200 Years of Syntax: A Critical Survey*. Amsterdam: John Benjamins.

Greenberg, Joseph H. (1966). "Some universals of grammar with particular reference to the order of meaningful elements". In: *Universals of language*. Ed. by Joseph H. Greenberg. Cambridge, Mass: MIT Press, pp. 73–113.

Gries, Stefan Th. (2001). "A Multifactorial Analysis of Syntactic Variation: Particle Movement Revisited". In: *Journal of Quantitative Linguistics* 8.1, pp. 33–50.

Gries, Stefan Th. (2009). *Quantitative Corpus Linguistics with R. A Practical Introduction*. New York: Routledge.

Groß, Thomas and Timothy Osborne (2015). "The Dependency Status of Function Words: Auxiliaries". In: *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*. Uppsala, Sweden: Uppsala University, Uppsala, Sweden, pp. 111–120.

Güldemann, Tom (2010). "The relation between focus and theticity in the Tuu family". In: *The Expression of Information Structure. A documentation of its diversity across Africa*. Ed. by Ines Fiedler and Anne Schwarz, pp. 69–93.

Hajič, Jan et al. (2004). "Prague Arabic dependency treebank: Development in data and tools". In: *In Proc. of the NEMLAR Intern. Conf. on Arabic Language Resources and Tools*, pp. 110–117.

Hale, Ken (1983). "Warlpiri and the Grammar of Non-Configurational Languages". In: *Natural Language & Linguistic Theory* 1.1, pp. 5–47.

Halliday, M.A.K. (2014). *Halliday's introduction to functional grammar (Fourth Edition)*. Ed. by Christian M.I.M. Matthiessen. Fourth Edition. London and New York: Routledge.

Haspelmath, Martin (2004). "Does linguistic explanation presuppose linguistic description?" In: *Studies in Language* 28.3, pp. 554–579.

— (2009). "Framework-Free Grammatical Theory". In: *The Oxford Handbook of Linguistic Analysis*. Ed. by Bernd Heine and Heiko Narrog. Oxford: Oxford University Press, pp. 341–365.

Hawkins, John A. (1983). *Word Order Universals*. New York: Academic Press.

Heine, Bernd and Heiko Narrog, eds. (2010). *The Oxford Handbook of Linguistic Analysis*. Oxford Handbooks in Linguistics. Oxford University Press.

Herbst, Thomas et al. (n.d.). *A Valency Dictionary of English: A Corpus-Based Analysis of the Complementation Patterns of English Verbs, Nouns and Adjectives*. 1484000/.

Heylen, Kris (2005). "A Quantitative Corpus Study of German Word Order Variation". In: *Linguistic evidence : empirical, theoretical, and computational perspectives*. Berlin / New York: Mouton de Gruyter, pp. 241–263.

Holmberg, Andres (2015). "Verb Second". In: *Syntax - Theory and Analysis: An International Hanbook. Volume 1*. Ed. by Tibor Kiss. Berlin: De Gruyter Mouton, pp. 342–383.

Hornstein, Norbert (2009). *A Theory of Syntax: Minimal Operations and Universal Grammar*.

Iggesen, Oliver A. (2013). "The World Atlas of Language Structures Online". In: *The World Atlas of Language Structures Online*. Ed. by Matthew S. Dryer and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology. Chap. Feature 49A: Number of Cases.

Itkonen, Esa (2005). "Concerning the Synthesis between Intuition-based Study of Norms and Observation-based Study of Corpora". In: *SKY Journal of Linguistics* 18, pp. 357–377.

Jackendoff, Ray (1977). *X-bar-Syntax: A Study of Phrase Structure*. Linguistic Inquiry Monograph 2. Cambridge, MA: MIT Press.

Jensen, Torben Juel and Tanya Karoli Christensen (2013). "Promoting the demoted: The distribution and semantics of 'main clause word order' in spoken Danish complement clauses". In: *Lingua* 137, pp. 38–58.

Johnson, Keith (2008). *Quantitative Methods in Linguistics*. Blackwell Publishing.

Jurafsky, Daniel and James H. Martin (2009). *Speech and Language Processing (2Nd Edition)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

Kallulli, Dalina and Liliane Tasmowski, eds. (2008). *Clitic doubling in the Balkan languages*. Amsterdam/Philadelphia: John Benjamins.

Kalmár, Ivan and Dionisius Agius (1981). "Verb-Subject Order in Maltese". In: *Journal of Maltese Studies* 14, pp. 22–31.

— (1983). "Verb-Subject Order and Communicative Dynamism in Maltese". In: *Zeitschrift für Dialektologie und Linguistik* 50.3, pp. 335–354.

Karimi, Simin, ed. (2003). *Word Order and Scrambling*. Malden, MA / Oxford: Blackwell Publishing. Chap. 10, pp. 217–237.

Katz, Jerrold J. (1996). "The Unfinished Chomskyan Revolution". In: *Mind and Language* 11.3, pp. 270–294.

Kepser, Stephan and Marga Reis, eds. (2005). *Linguistic evidence : empirical, theoretical, and computational perspectives*. Berlin / New York: Mouton de Gruyter.

Kerstens, J.G. (1975). *Over afgeleide structuur en de interpretatie van zinnen*.

Kiss, Katalin É., ed. (1995a). *Discourse Configurational Languages*. Oxford: Oxford University Press.

— (1995b). "Discourse Configurational Languages: Introduction". In: *Discourse Configurational Languages*. Ed. by Katalin É. Kiss. Oxford: Oxford University Press, pp. 3–27.

Kiss, Tibor and Artemis Alexiadou, eds. (2015). *Syntax - Theory and Analysis: An International Hanbook. Volume 1*. Berlin: De Gruyter Mouton.

Köhler, Reinhard (2012). *Quantitative Syntax Analysis*. Berlin/Boston: De Gruyter Mouton.

Krapova, Iliyana and Guglielmo Cinque (2008). "Clitic reduplication constructions in Bulgarian". In: *Clitic doubling in the Balkan languages*. Ed. by Dalina Kallulli and Liliane Tasmowski. Amsterdam/Philadelphia: John Benjamins, pp. 257–287.

Krause, Thomas and Amir Zeldes (2016). "ANNIS3: A new architecture for generic corpus query and visualization". In: *Digital Scholarship in the Humanities* 31.1, pp. 118–139.

Krejčová, Elena (2016). *Slovosledné změny v bulharských a srbských evangelních památkách z 12. a 13. století*. Brno: Filozofická fakulta, Masarykova Univerzita.

Krier, Fernande (1976). *Le maltais au contact de l'italien. Etude phonologique, grammatical et sémantique*. Hamburg: Helmut Buske Verlag.

Krifka, Manfred (2007). "Basic notions of information structure". In: *Working Papers of the SFB 632. Interdisciplinary Studies on Information Structure*. Ed. by Caroline Féry, Gisbert Fanselow, and Manfred Krifka. Potsdam: Universitätsverlag Potsdam, pp. 13–55.

Krifka, Manfred and Renate Musan (2012). "Information structure: Overview and linguistic issues". In: *The Expression of Information Structure*. Ed. by Manfred Krifka and Renate Musan. Berlin/Boston: De Gruyter Mouton, pp. 1–44.

Kuroda, S.-Y. (1972). "The Categorical and the Thetic Judgment: Evidence from Japanese Syntax". In: *Foundations of Language* 9 (2).

Lahdo, Ablahad (2009). *The Arabic Dialect of Tillo in the Region of Siirt (South-eastern Turkey). Acta Universitatis Upsaliensis*. Studia Semitica Upsaliensia 26. Uppsala: Uppsala University.

Lehmann, Winfred P., ed. (1978). *Syntactic Typology: Studies in the Phenomenology of Language*. Austin: University of Texas Press.

Levinson, Stephen C. (2008). *Pragmatics*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig (2016). *Ethnologue: Languages of the World, Nineteenth edition*. Dallas, Texas: SIL International.

Lieber, Rochelle and Pavol Štekauer (2009). "Introduction: Status and Definition of Compounding". In: *The Oxford Handbook of Compounding*. Ed. by Rochelle Lieber and Pavol Štekauer. Oxford: Oxford University Press, pp. 3–18.

Lopatková, Markéta et al. (2017). *Valenční slovní českých sloves VALLEX*. Praha: Karolinum.

Lüdecke, Daniel (2017). *sjPlot: Data Visualization for Statistics in Social Science*. R package version 2.4.0.

Maas, Utz (2009). "Complex predicates in Maltese: From a Neo-Arabic perspective". In: *Introducing Maltese Linguistics*. Ed. by Bernard Comrie et al. Amsterdam: John Benjamins, pp. 113–132.

Mahajan, Anoop (2003). "Word Order and (Remnant) VP Movement". In: *Word Order and Scrambling*. Ed. by Simin Karimi. Malden, MA / Oxford: Blackwell Publishing. Chap. 10, pp. 217–237.

Maiden, Martin and Cecilia Robustelli (2007). *A Reference Grammar of Modern Italian*. London and New York: Routledge.

Manfredi, Stefano (2017). "The encoding of pain in modern Arabic dialects: a typological perspective". A paper read at the 12th International Conference on Arabic Dialectology AIDA, Université Aix-Marseille, May 30 – June 2, 2017.

Martinet, André (1969). *Langue et fonction*. Paris: Gonthier/Denoël.

— (1970). *Eléments de linguistique générale*. Paris: Colon.

Marty, Anton (1897). "Über die Scheidung von grammatischem, logischem und psychologischem Subjekt resp. Prädikat". In: *Archiv für Systematische Philosophie* 3, pp. 174–189.

Mathesius, Vilém (1907). "Studie k dějinám anglického slovosledu. 1. Kapitoly úvodní". In: *Věstník České akademie císaře Františka Josefa pro vědy, slověsnost a umění* XVI, pp. 261–274.

— (1939). "O tak zvaném aktuálním členění věty". In: *Slovo a slovesnost* 5, pp. 171–174.

— (1961). *Obsahový rozbor současné angličtiny na základě obecně lingvistickém*. Ed. by Josef Vachek. Praha: Nakladatelství Československé akademie věd.

Matthews, P. H. (1981). *Syntax*. Cambridge: Cambridge University Press.

McCawley, James D. (1970). "English as a VSO Language". In: *Language* 46.2, pp. 286–299.

McDonald, Ryan et al. (2013). "Universal Dependency Annotation for Multilingual Parsing". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 92–97.

McNally, Louise (2011). "Existential sentences". In: *Semantics: an international hanbook of natural language meaning*. Ed. by Klaus von Heusinge, Claudia Maienborn, and Paul Portner. Vol. 2. Handbooks of linguistics and communication science; 33.2. Berlin/Boston: De Gruyter Mouton, pp. 1829–1848.

Meisel, Jürgen M. and Martin D. Pam, eds. (1979). *Linear Order and Generative Theory*. Current Issues in Linguistic Theory 7. John Benjamins B. V.

Meurman-Solin, Anneli, María José López-Couso, and Bettelou Los, eds. (2012). *Information Structure and Syntactic Change in the History of English*. Oxford: Oxford University Press.

Mifsud, Manwel (1995). *Loan Verbs in Maltese: A Descriptive and Comparative Study*. Leiden: Brill.

Miller, Jim (1995). "Does spoken language have sentences?" In: *Grammar and meaning: Essays in honour of Sir John Lyons*. Ed. by F. R. Palmer. Cambridge: Cambridge University Press, pp. 116–135.

Nakagawa, Natsuko (2018). *Information structure in spoken Japanese: Particles, word order, and intonation*. Berlin: Language Science Press.

Newmeyer, Frederick J. (2005). *Possible and and probable languages: a generative perspective on linguistic typology*. Oxford: Oxford University Press.

Nikanne, Urpo (2017). "Finite sentences in Finnish: Word order, morphology, and information structure". In: *Order and structure in syntax I: Word order and syntactic structure*. Ed. by Laura R. Bailey and Michelle Sheehan. Vol. I. Berlin: Language Science Press, pp. 29–47.

Nivre, Joakim, Željko Agić, et al. (2017). *Universal Dependencies 2.1*. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Nivre, Joakim, Filip Ginter, et al. (2014). *Online documentation for Universal Dependencies, version 1 (2014-10-01)*.

— (2016). *Online documentation for Universal Dependencies, version 2 (2016-12-01)*.

Nivre, Joakim and Eva Hajičová, eds. (2015). *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*. Uppsala, Sweden: Uppsala University, Uppsala, Sweden.

Nivre, Joakim, Marie-Catherine de Marneffe, et al. (2016). "Universal Dependencies v1: A Multilingual Treebank Collection". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Ed. by Nicoletta Calzolari (Conference Chair) et al. Portorož, Slovenia: European Language Resources Association (ELRA).

Olsen, Mari B. (2014). *A Semantic and Pragmatic Model of Lexical and Grammatical Aspect*. New York: Routledge.

Osborne, Timothy and Daniel Maxwell (2015). "A Historical Overview of the Status of Function Words in Dependency Grammar". In: *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*. Uppsala, Sweden: Uppsala University, Uppsala, Sweden, pp. 241–250.

Osborne, Timothy, Michael Putnam, and Thomas Groß (2012). "Catenae: Introducing a Novel Unit of Syntactic Analysis." In: *Syntax* 15.4, pp. 354–396.

Owens, Jonathan and Alaa Elgibali, eds. (2010). *Information Structure in Spoken Arabic*. London: Routledge.

Paggio, Patrizia and Alexandra Vella (2014). "Overlaps in Maltese Conversational and Task Oriented Dialogues". In: *Proceedings of the 1th European Symposium on Multimodal Communication*. Linköping Electronic Conference Proceedings.

Panevová, Jarmila et al. (2014). *Mluvnice současné češtiny. 2, Syntax na základe anotovaného korpusu*. Praha: Karolinum.

Panhuis, Dirk G.J. (1982). *The Communicative Perspective in the Sentence: A Study of Latin Word Order*. Amsterdam/Philadelphia: John Benjamins.

Panzavecchia, Fortunato (1845). *Grammatica della lingua maltese spiegata secondo il principj delle lingue orientali e della lingua italiana*. Malta: Tipografia di M. Weiss.

Payne, Thomas E. (1997). *Describing Morphosyntax: A Guide for Field Linguists*. Cambridge: Cambridge University Press.

Perini, Mário A. (2015). *Describing Verb Valency. Practical and Theoretical Issues*. Heidelberg: Springer.

Peterson, John (2009). "'Pseudo-verbs': An analysis of non-verbal (co-)predication in Maltese". In: *Introducing Maltese Linguistics*. Ed. by Bernard Comrie et al. Amsterdam: John Benjamins, pp. 181–206.

Petrov, Slav, Dipanjan Das, and Ryan McDonald (2012). "A Universal Part-of-Speech Tagset". In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Ed. by Nicoletta Calzolari (Conference Chair) et al. Istanbul, Turkey: European Language Resources Association (ELRA).

Petrova, Svetlana and Michael Solf (2008). "Rhetorical relations and verb placement in the early Germanic languages: A cross-linguistic study". In: *'Subordination' versus 'Coordination' in Sentence and Text*. Ed. by Cathrine Fabricius-Hansen and Wiebke Ramm. Trends in Linguistics. Studies and Monographs [TiLSM] 227. Amsterdam / Philadelphia: John Benjamins, pp. 329–351.

Poole, Geoffrey (2017). "Feature inheritance in Old Spanish: (re)visiting V2". In: *Order and structure in syntax I: Word order and syntactic structure*. Ed. by Laura R. Bailey and Michelle Sheehan. Vol. I. Berlin: Language Science Press, pp. 49–68.

Pullum, Geoffrey K. (2017). "Theory, data, and the epistemology of syntax". In: *Proceedings of the 2016 conference of the Institut für Deutsche Sprache*.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Ritt-Benmimoun, Veronika (2014). *Grammatik des arabischen Beduinendialekts der Region Douz (Südtunesien)*. Wiesbaden: Harrassowitz.

Rizzi, Luigi (1997). "The Fine Structure of the Left Periphery". In: *Elements of Grammar: Handbook of Generative Syntax*. Dordrecht: Kluwer Academic Publishers, pp. 281–337.

— ed. (2004). *The Structure of CP and IP: The Cartography of Syntactic Structures Volume 2*. 1175000/: Oxford University Press.

Roberts, Ian G. (2005). *Principles and Parameters in a VSO Language: A Case Study in Welsh*. Oxford Studies in Comparative Syntax. Oxford: Oxford University Press.

Roby, David Brian (2009). *Aspect and the Categorization of States: The Case of Ser and Estar in Spanish*. Studies in Language Companion Series, 114. Amsterdam: John Benjamins.

Rosengren, Inger (1997). "The thetic/categorical distinction revisited once more". In: *Linguistics* 35, pp. 439–479.

Rosner, Mike, Joe Caruana, and Ray Fabri (2000). *Maltilex: A Computational Lexicon for Maltese*. Msida: University of Malta.

Ross, John R. (1967). "Constraints on variables in syntax". PhD thesis. Massachusetts Institute of Technology.

Roudanovsky, Basil (1910). *Maltese Pocket Grammar*. Valletta: John Critien.

— (1911). *Quelques particularités de dialecte arabe de Malte*. 2me édition, revue et augmentée. Beyrouth: Imprimerie catholique.

Rubio, Gonzalo (2009). "Semitic Influence in the History of Latin Syntax". In: *New Perspectives on Historical Latin Syntax. Volume 1: Syntax of the Sentence*. Ed. by Philip Baldi and Pierluigi Cuzzolin. Mouton de Gruyter. Chap. Semitic influence in the history of Latin syntax, pp. 195–239.

Rychlý, Pavel (2007). "Manatee/Bonito - A Modular Corpus Manager". In: *1st Workshop on Recent Advances in Slavonic Natural Language Processing*. Brno: Masaryk University, pp. 65–70.

Rysová, Kateřina and Jiří Mírovský (2014). "Valency and Word Order in Czech - A Corpus Probe". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Ed. by Nicoletta Calzolari (Conference Chair) et al. Reykjavik, Iceland: European Language Resources Association (ELRA).

Saade, Benjamin and Mauro Tosco, eds. (2017). *Advances in Maltese Linguistics*. Berlin: De Gruyter.

Sasse, Hans-Jürgen (1987). "The thetic/categorical distinction revisited". In: *Linguistics* 25, pp. 511–580.

— (1995). "'Theticity' and VS order". In: *Sprachtypol, Univ. Forsch. (STUF)* 48.1/2, pp. 3–31.

Saussure, Ferdinand de (1995). *Course de linquistique générale*. Ed. by Charles Bailly et al. Paris: Éditions Payot & Rivages.

Savary, Agata et al. (2018). "PARSEME multilingual corpus of verbal multiword expressions". In: *Representation and Parsing of Multiword Expressions*. Ed. by Yannick Parmentier and Jakub Waszczuk. Berlin: Language Science Press.

Al-Sayyed, Amany and David Wilmsen (2017). "Verbal negation with muš in Maltese and Eastern Mediterranean Arabics". In: *Advances in Maltese Linguistics*. Ed. by Benjamin Saade and Mauro Tosco. Berlin: De Gruyter, pp. 151–172.

Sgall, Petr (1967a). "Functional Sentence Perspective in a Generative Description". In: *Prague Studies in Mathematical Linguistics* 2, pp. 203–225.

— (1967b). "K formálním vlastnostem syntaktických vztahů". In: *Slovo a slovesnost* 28.4, pp. 359–363.

Sgall, Petr, Allevtina Bémová, et al. (1986). *Úvod do syntaxe a sémantiky: některé nové směry v teoretické lingvistice*. Praha: Academia.

Sgall, Petr, Eva Hajičová, and Eva Buráňová (1980). *Aktuální členění věty v češtine*. Praha: Academia.

Siewierska, Anna (1988). *Word Order Rules*. Croom Helm linguistics series. London/New York/Sydney: Croom Helm.

— ed. (1998). *Eurotyp: Typology of Languages in Europe, Volume 1: Constituent Order in the Languages of Europe*. Empirical Approaches to Language Typology 20.1. Mouton de Gruyter.

Silveira, Natalia et al. (2014). "A Gold Standard Dependency Corpus for English". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Ed. by Nicoletta Calzolari (Conference Chair) et al. Reykjavik, Iceland: European Language Resources Association (ELRA).

Sitaridou, Ioanna (2011). "Word order and information structure in Old Spanish". In: *Catalan Journal of Linguistics* 10.0, pp. 159–184.

Skënderi, Arion (1997). *Functional Sentence Perspective in Albanian Texts*. Dissertationen der Universität Wien 41. Wien: WUV-Universitätsverlag.

Song, Jae Jung, ed. (2011a). *The Oxford Handbook of Linguistic Typology*. Oxford: Oxford University Press.

— (2011b). "Word Order Typology". In: *The Oxford Handbook of Linguistic Typology*. Ed. by Jae Jung Song. Oxford: Oxford University Press, pp. 253–279.

Sornicola, Rosanna (1994). "On Word-Order Variability: A Study from a Corpus of Italian". In: *Lingua e stile* XXIX.1, pp. 25–57.

Souag, Lameen (2017). "Clitic Doubling and Contact in Arabic". In: *Zeitschrift für Arabische Linguistik* 66, pp. 45–70.

Spagnol, Michael (2011). "A Tale of Two Morphologies. Verb structure and argument alternations in Maltese". PhD thesis. Universität Konstanz.

Stenetorp, Pontus et al. (2012). "BRAT: A Web-based Tool for NLP-assisted Text Annotation". In: *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. EACL '12. Avignon, France: Association for Computational Linguistics, pp. 102–107.

Stolz, Thomas (2009). "Splitting the verb chain in modern literary Maltese". In: *Introducing Maltese Linguistics*. Ed. by Bernard Comrie et al. Amsterdam: John Benjamins, pp. 133–179.

— (2011). "The possessive relative clause in Maltese". In: *Variation and Change. The Dynamics of Maltese in Space, Time and Society*. Ed. by Sandro Caruana, Ray Fabri, and Thomas Stolz. Studia Typologica 9. Berlin: Akademie Verlag, pp. 183–232.

Stolz, Thomas and Nataliya Levkovych (2017). "From variation towards the grammar of Maltese prepositions – first steps". Paper presented at the Sixth International Conference on Maltese Linguistics, June 8th – June 9th, Comenius University in Bratislava.

Stolz, Thomas and Benjamin Saade (2016). "On short and long forms of personal pronouns in Maltese". In: *Shifts and Patterns in Maltese*. Ed. by Gilbert Puech and Benjamin Saade. Berlin/Boston: Walter de Gruyter, pp. 199–268.

Stumme, Hans (1904). *Maltesische Studien: Eine Sammlung prosaischer und poetischer Texte in maltesischer Sprache, nebst Erläuterungen*. Leipzig: Heinrichs.

Sutcliffe, Edmund (1936). *A Grammar of the Maltese Language. With Chrestomathy and Vocabulary*. Oxford: Oxford University Press.

Suter, Hans-Jürg (1993). *The Wedding Report. A prototypical approach to the study of traditional text types*. Amsterdam: John Benjamins.

Tagħrif fuq il-Kitba Maltija (1924).

Taylor, Ann and Susan Pintzuk (2012). "The Effect of Information Structure on Object Position in Old English: A Pilot Study". In: *Information Structure and Syntactic Change in the History of English*. Ed. by Anneli Meurman-Solin, María José López-Couso, and Bettelou Los. Oxford: Oxford University Press, pp. 47–65.

Tesnière, Lucien (1959). *Éléments de syntaxe structurale*. Paris: Librairie C. Klincksieck.

— (2015). *Elements of structural syntax*. Amsterdam: John Benjamins.

Thuilier, Juliette, Anne Abeillé, and Benoît Crabbé (2014). "Ordering preferences for postverbal complements in French". In: *French through Corpora - Ecological and Data-Driven Perspectives in French Language Studies*. Ed. by Henry Tyne et al. Newcastle upon Tyne: Cambridge Scholars Publishing.

Tonhauser, Judith and Erika Colijn (2010). "Word Order In Paraguayan Guaraní". In: *International Journal of American Linguistics* 76.2, pp. 255–288. eprint: https://doi.org/10.1086/652267.

Tyne, Henry et al., eds. (2014). *French through Corpora - Ecological and Data-Driven Perspectives in French Language Studies*. Newcastle upon Tyne: Cambridge Scholars Publishing.

Uhlířová, Ludmila (1967). "Statistics of the Word Order of Direct Object in Czech". In: *Prague Studies in Mathematical Linguistics* 2, pp. 37–49.

Ussishkin, Adam, Jerid Francom, and Dainon Woudstra (2009). "Creating a web-based lexical corpus and information-extraction tools for the Semitic language Maltese". In: *Information Retrieval and Information Extraction for Less Resourced Languages: IE-IR-LRL, Donostia, September 7th, 2009. Proceedings*. Ed. by Iñaki Alegria, Mikel L. Forcada, and Kepa Sarasola. University of the Basque Country. SALTMIL, pp. 9–16.

Vaculíková, Petra and Michal Jurka et al., eds. (2015). *Tematicko-rematický nexus z rozmanitých perspektiv v různých jazycích*. Olomouc: Univerzita Palackého v Olomouci.

Vallduví, Enric and Elisabet Engdahl (1996). "The linguistic realization of information packaging". In: *Linguistics* 34, pp. 459–519.

Vanhove, Martine (1993). *La langue maltaise. Etudes syntaxiques d'un dialecte arabe »périphérique«*. Wiesbaden: Harrassowitz Verlag.

Vassalli, Michelantonio (1791). *Mylsen phoenico-punicum sive grammatica melitensis*. Roma.

— (1827). *Grammatica della lingua maltese*. Malta: Self-published.

Vella, Francis (1831). *Maltese Grammar for the Use of the English*. Leghorn: Glaucus Masi.

Vella, Joseph (1970). *A Comparative Study in Maltese and Libyan (Benghazi Dialect). Phonetics, Morphology, Syntax & Lexicon*. Malta: The Royal University of Malta (Unpublished PhD thesis).

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S*. Fourth. ISBN 0-387-95457-0. New York: Springer.

Vennemann, Theo (1974). "Topics, subjects, and word order: From SXV to SVX via TVX". In: *Historical Linguistics: Proceedings of the First International Congress of Historical Linguistics, Edinburgh, September 1973*. Ed. by John Anderson and Charles Jones. Vol. II. Amsterdam (North-Holland), pp. 339–376.

Villalba, Xavier (2000). "The Syntax of Sentence Periphery". PhD thesis. Universitat Autònoma de Barcelona.

— (2011). "A quantitative comparative study of right-dislocation in Catalan and Spanish". In: *Journal of Pragmatics* 43.7, pp. 1946–1961.

Vincent, Nigel (1979). "Word Order and Grammatical Theory". In: *Linear Order and Generative Theory*. Ed. by Jürgen M. Meisel and Martin D. Pam. Current Issues in Linguistic Theory 7. Amsterdam: John Benjamins B. V.

von der Gabelentz, Georg (1869). "Ideen zu einer vergleichenden Syntax". In: *Zeitschrift für Völkerpsychologie und Sprachwissenschaft* 6, pp. 376–384.

Wahrmund, Adolf (1861). *Praktisches Handbuch der neu-arabischen Sprache*. Giefsen: J. Ricker'sche Buchhandlung.

Wallis, Sean (2013). "Binomial Confidence Intervals and Contingency Tests: Mathematical Fundamentals and the Evaluation of Alternative Methods". In: *Journal of Quantitative Linguistics* 20.3, pp. 178–208.

Wasow, Thomas and Jennifer Arnold (2003). "Post-verbal constituent ordering in English". In: *In Rohdenburg G. and Mondorf B. (Eds), Determinants of Grammatical Variation in English*, pp. 119–154.

Weil, Henri (1844). *De l'ordre des mots dans les langues anciennes comparées aux langues modernes*. Paris: Joubert.

Welke, Klaus (2011). *Valenzgrammatik des Deutschen. Eine Einführung*. Berlin / Boston: De Gruyter.

Wzzino, Francesco (1752). *Taġhlim Nisrani*. Roma.

Yimam, Seid Muhie et al. (2014). "Automatic Annotation Suggestions and Custom Annotation Layers in WebAnno". In: *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Baltimore, Maryland: Association for Computational Linguistics, pp. 91–96.

Zanuttini, Raffaella et al., eds. (2006). *Crosslinguistic Research in Syntax and Semantics: Negation, Tense, and Clausal Architecture*. Washington, D.C.: Georgetown University Press.

Zeman, Daniel (2008). "Reusable Tagset Conversion Using Tagset Drivers". In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Ed. by Nicoletta Calzolari (Conference Chair) et al. http://www.lrec-conf.org/proceedings/lrec2008/. Marrakech, Morocco: European Language Resources Association (ELRA).

Zeman, Daniel et al. (2014). "HamleDT: Harmonized Multi-Language Dependency Treebank". In: *Language Resources and Evaluation* 48.4, pp. 601–637.

Zifonoun, Gisela (2003). "Grundlagen der Valenz". In: *Dependenz und Valenz. Ein internationales Handbuch der zeitgenössischen Forschung / Dependency and Valency. An International Handbook of Contemporary Research*. Ed. by Vilmos Ágel et al. Vol. 1. Berlin / New York: Walter de Gruyter, pp. 352–377.

Zikánová, Šárka (2009). *Postavení slovesného přísudku ve starší češtině (1500 - 1620)*. Praha: Karolinum.

Zimmermann, Malte (2016). "Predicate Focus". In: *The Oxford Handbook of Information Structure*. Ed. by Caroline Féry and Shinichiro Ishihara. Oxford: Oxford University Press, pp. 314–335.

Zipser, Florian and Laurent Romary (2010). "A model oriented approach to the mapping of annotation formats using standards." In: *Workshop on Language Resource and Language Technology Standards, LREC 2010*. La Valette, Malta.

Zubizarreta, Maria Luisa (1998). *Prosody, Focus, and Word Order*. Cambridge, Massachusetts and London, England: The MIT Press.

Zwart, Jan-Wouter (2017). "An argument against the syntactic nature of verb movement". In: *Order and structure in syntax I: Word order and syntactic structure*. Ed. by Laura R. Bailey and Michelle Sheehan. Vol. I. Berlin: Language Science Press, pp. 69–97.

Abbreviations

This list contains abbreviations and acronyms other than those used for part-of-speech tags (Chapter 5, section 5.4.1), UD annotation levels (Chapter 6, section 6.3), UD relations (Chapter 6, section 6.4.3) and VALLEX actants and free dependents (Chapter 6, section 6.4.4.2.2).

A agent in passive clauses
ACT active
ACC accusative
BCv3 bulbulistan corpus malti v3
CA Classical Arabic
COMP complementizer
COP copula
DAT dative
DEF definite article
EL Greek
EN English
EXIST existential predicate
F feminine
FGP Functional Generative Description
FSP Functional Sentence Perspective
FUT future marker
GEN genitive
FUT future marker
HU Hungarian
HUUDv2 Hungarian Universal Dependencies Treebank version 2
I indirect object
ID identifier
INT interjection
INTR interrogative suffix
LF Logical Form
LFG Lexical Functional Grammar
M masculine
MLRSv3 Maltese Language Resource Server, *Korpus Malti v3.0*
MP The Minimalist Program
MT Maltese
MUDTv1 Maltese Universal Dependencies Treebank version 1 (current version)
MUDTv2 Maltese Universal Dependencies Treebank version 2 (future version)
NA Neo-Arabic
NEG negation

O object
P&P Principles and Parameters
PF Phonetic Form
PART participle
PASS passive
PADT Prague Arabic Dependency Treebank
PDT Prague Dependency Treebank
PL plural
PROG progressive marker
PTG XML-based file format for PosTagger
Rh rheme
S subject
SG singular
Th theme
Tr transition
UD Universal Dependencies
UG Universal Grammar
V verbal, copular or existential predicate
V2 verb-second languages
v version
VOC vocative
WALS The World Atlas of Language Structures

Appendix

BCv3 (Chapter 5) can be accessed at bulbul.sk/bonito2/ (login name: guest, password: Ghilm3); MUDTv1 (Chapter 6) is accessible as an ANNIS3 instance at bulbul.sk/annis-gui-3.4.4 and as a set of HTML files at bulbul.sk/bonito2/treebank (login name: guest, password: Ghilm3).

The source files for these corpora and other supplementary information (lists of works in *BCv3*, data and code used in Chapter 7 etc.) are provided in the appendices. These can be accessed online at http://bulbul.sk/phd and obtained on a DVD by writing to the author at bulbul@bulbul.sk. The contents of the appendices are described below.

Appendix A contains detailed descriptions of

1. text subtypes for text type newspaper,
2. sources for text type fiction, and
3. sources for text type non-fiction

stored in the directory AppendixA in the following file structure:

```
AppendixA
├── newspaper
├── fiction
└── non-fiction
```

Appendix B can be found in the AppendixB directory and contains the raw data for *BCv3* and MUDTv1 in the following subdirectories:

```
AppendixB
├── BCv3
│   ├── Registry_files
│   └── Compiled
└── MUDTv1
    ├── PTG
    └── CoNLL-U
```

Appendix C contains the list of ANNIS queries employed to retrieve the data used in Chapter 7, sections 7.2, 7.3 and 7.4 stored in the subdirectory AppendixC. Additionally, it contains the R scripts and CSV source files used to create calculations, plots and tables in Chapter 7, sections 7.2, 7.3 and 7.4. The structure of the files is as follows: