**REPORT ON PH.D. THESIS OF MR. SLAVOMÍR ČEPLÖ**

**<u>28th May 2018</u>**

**Student:** Slavomír Čéplö

**Title:** Constituent order in Maltese: A quantitative analysis

**Examiner:** Prof. Ray Fabri, Institute of Linguistics and Language Technology, University of Malta

**Report:**

Mr Čéplö's research is a quantitative, corpus based study of constituent order in Maltese at clause level. Basing his analysis on a written corpus, he specifically sets out to determine the dominant constituent order, as well as the degree and reasons for deviation and variation from the dominant order.

The dissertation comprises seven main chapters and a final section with a summary/outlook, as well as a list of references, abbreviations and appendices available online. The main chapters cover the author's approach to research in linguistics, a review of the relevant literature on constituent order in general and in Maltese, research questions and methodology, the corpus of Maltese used as a data source, a sketch of Maltese syntax and the Maltese Treebank, a quantitative analysis of constituent order based on the Treebank, and the conclusions reached on the basis of the analysis carried out. Mr Čéplö concludes that Maltese cannot be said to be have 'free word order', or to be 'discourse-configurational' or 'topic-prominent', as has been claimed previously in a number of studies.

## Chapter 1

In chapter one, Čéplö starts by clarifying the approach he takes with respect to the 'nature and goals' of linguistics, describing his approach as being 'descriptive' and 'empirical' (p. 1). Moreover, his description of features of the grammar Maltese, the language under focus, lies 'outside of any existing theoretical framework…on its own, without any conscious preconceptions or biases' (p. 1). Čéplö very strongly argues in favour of his own empirical, i.e., essentially corpus based, approach, as opposed to other 'introspective or intuitive approaches' (p. 2). This is, of course, a recurrent and hotly debated topic in language research, in particular since the beginnings of generative linguistics in the 1950's with Chomsky's critique of corpora and focus on introspection. The debate is still very topical, especially since the relative recent rapid development of digital corpora and so-called big data as rich sources of information.

Čéplö makes a good case for the approach he has opted for, and is consistent in applying it throughout the study. Indeed, Čéplö himself relativises his strong stance in favour of corpus based approaches when he claims that 'Experience has shown that corpus data, however large, may not be sufficient...Any full description of any language should thus make full advantage of all data collection tools available to a linguist, including elicitation and experimentation' (p. 2).

I cannot but agree with this stance: all approaches have their pros and cons, whether inductive or deductive, framed within a specific theoretical framework or 'theory-free' (assuming that is at all possible), or based on different data sources. Ideally analyses are based on data obtained from various sources (corpora, elicitation, etc.). However, it is not only practical but also desirable for the researcher to choose an approach and stick to it as far as possible, and this is what Čéplö does consistently but also self-critically throughout his study. Indeed, this is typical of Čéplö's general attitude: he carefully looks at all aspects related to an issue in a comprehensive and balanced manner, and then justifies and is consistent in applying his own preference. This is definitely one of the strong points in this study.

One minor criticism I would like to mention here is that Čéplö occasionally hints at arguments but leaves one to guess what they are. For example, on page 2, he says 'I find this framing [that phenomenological description entails accurate prediction of speaker behavior] troubling for several reasons', but then goes on to give only one reason ('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'). If one feels so strongly about a matter, then one should make a robust case for it and not merely hint at the 'several' reasons for it.

In this chapter, Čéplö also goes to great pains to clarify his use of terminology, especially since some terms are often taking for granted in the literature. These include terms for basic concepts, as can be seen from the subsections dedicated to each term: Maltese, Dependency Syntax, Sentence, Word, Predicate, Clause, Phrase, Constituent, and Pragmatics. The discussion is always clear and plausible, with Čéplö identifying and ironing out any potential inconsistencies in the way he adopts a term, as when he discusses the concept of 'syntax': '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' (p4). Basic terms which are often used imprecisely are clarified, such as when Čéplö clearly states that, in his study, 'constituent order' refers to 'the order in which the predicate and its core arguments…appear in a sentence', while 'word order' refers to 'the order of elements with [sic] a phrase (such as the order of nouns and adjectives or adjectives and adverbs)' (p. 8).

Of particular importance is the clarification about the language itself being analysed, i.e., Maltese, or, to be exact, current ('first two decades of the 21st century') written Maltese. The limitation to the written language, as opposed to spoken Maltese, is justified and is the result of the lack of spoken corpora of Maltese as opposed to the availability of written corpora. This is crucial because, at least intuitively, there are bound to be significant differences between the spoken and written mediums, in particular in the possible effects that discourse factors (deixis, topicality, discourse participants, flow of information, etc.) can have on constituent order. Of course, the real extent of such differences, if any, can only be gauged once spoken data are available. Note that, in his chapter seven, Čéplö does consider the statistical possibility of projecting his conclusions about constituent order in the written medium to be the spoken medium, and speculates that there are indications that his main conclusion will probably still hold when applied to spoken data.

**Chapters 2 and 3**

Chapters 2 and 3 review the literature on theories of and approaches to constituent order, in general (chapter 2), and on studies on constituent order in Maltese, in particular (chapter 3). Chapter 2 gives an overview of studies on constituent order, ranging from the classic typological work by Greenberg to generative approaches (Aspects to the Minimalist Program), also including functional perspectives and quantitative (corpus based) approaches. Čéplö covers the main approaches to the study of constituent order and gives a concise and critical overview of this rich field of study. Perhaps the overview could have gained more in terms of thoroughness if the focus on generative approaches (sections 2.3.1 – 2.3.3) were not limited to the 'Chomskian' approach, and at least also mentioned non-Chomskian generative approaches, such as Lexical Functional Grammar, Head-Driven Phrase Structure Grammar and Optimality Theory.

Chapter 3 also presents a thorough and detailed overview of studies that either specifically deal with constituent order in Maltese, or refer to it in some manner. Čéplö critically reviews and evaluates the ideas and approaches taken by various linguists and grammarians who have contributed to the question of constituent order in Maltese. This chapter is very systematic and insightful, and provides a complete picture of the work that has been done in this area. He concludes that 'most such studies have been introspective at best, impressionistic at worst' (p. 49), and plans to remedy the situation with his study.

**Chapter 4**

After listing four research questions, Čéplö discusses each in turn, justifying each and, again, clarifying his use of terminology, in particular, the term referring to the concept of dominant constituent order, which is what the study essentially about. Čéplö uses Dryer's definition as a working definition for his study, i.e., an order is considered dominant 'if text counts reveal one order of a pair of elements to be more than twice as common as the other order' (p. 52). This definition fits in well with the corpus-based approach adopted by Čéplö, and he assumes since it serves his purposes. Any other non-dominant order is termed 'deviant' by Čéplö, who also clarifies the difference between 'variation' and 'deviation', both of which feature in the research questions.

It is a pity that Čéplö does not question Dryer's definition. Is it straightforward that that the dominant order is the one that is 'twice as common', and not say one and half times, or three times, as common. I would also question whether dominance should be equated with frequency in a corpus. Moreover, it would be in order to consider, and to contrast, the concept of 'dominant order with that of 'basic order', a term often used in the literature.

**Chapter 5**

Chapter 5 describes and discusses the history of digital corpora Maltese, and, in particular, the *bulbulistan maltiV3* (BCv3) corpus, which Čéplö himself developed, and which serves as the primary data source, for the study. Čéplö discusses in detail the processing and annotation (conversion, cleaning, splitting, tokenization, querying, tagging) of the data in preparation for the Treebank for Maltese, which is dealt with in chapter 6.

Čéplö is clear about the fact that both MLRSv3 and BCv3 are 'opportunistic corpora by nature' (p. 58) but, in order to limit the degree of opportunism, also places certain limitations on the texts included in BCv3', namely, that they be limited to written texts, original, publicly available and recent. In principle, non-representativeness of the corpus and the restriction to a written corpus is a problematic issue, but Čéplö is right in not being dissuaded by this drawback from carrying on with his analysis, and clearly very aware of and honest about these drawbacks. Indeed, the results thus obtained can eventually be tested when more representative written and, in particular, a spoken corpora are available.

The tagset used is also discussed in great detail, with examples, and one can argue on a number of decisions. The tags are aptly chosen, adequate and practical for the purpose at hand, although one might always argue about specific decisions. For example, English nouns modifying other English nouns (in compounds) are tagged as ADJ. Apart from whether such elements should be considered to be adjectives (also in English) rather than nouns, it is not clear why, in the example given (*kellna ħames gas/ADJ turbines/NOUN…* 'we had many gas turbines'), they are not tagged as X_ENG, since, in the discussion on English loan words, it is stated that the label X_ENG is applied if a 'sequence of tokens displays English syntax' (p 79). This is very much the case here since 'proper' adjectives in Maltese are post-nominal, not pre-nominal and, moreover, unlike English, modifier nouns in compounds in Maltese also occur to the right of the head noun (e.g., *linji gwida* 'guidelines' lit. 'lines guide'). The same logic applies to the example *Dawn huma d-double/ADJ standards/NOUN* (p. 79) 'These are the double standards'.

Another case is when the form *imsiefer* 'be abroad' (p. 74) is labelled as active participle, when, in fact, it is a passive participle, as shown by the 'm' prefix applied to the third form *siefer* 'go abroad' (just like *mbierek* 'blessed' and *bierek* 'bless'). A final example is on page 74 (section 5.4.1.3.22), where the label NUM_WHD is given to 'the word *wieħed* "one", its feminine form *waħda* and its plural *uħud'*. It is not clear whether a distinction is made between the numeral meaning of *wieħed* 'one' and (possibly rarer occurrence of) *wieħed* meaning 'a (certain)' as in *Kien hemm wieħed raġel* 'There was a (certain) man' as opposed to *Kien hemm raġel wieħed* 'There was one man', with a different distribution.

Finally, it is surprising that cases like *ssepara* 'separate', *sseparat* 'separated' and phrases *jiġi separat* "(it) is separated" turn out to be 'rare' in the corpus. This is a case, where native speaker intuitions based on everyday use seem to be in conflict with the objective corpus-based generalisations. It also raises the question of whether it is actual frequency in a recorded corpus that is relevant for processing, or, rather, whether it is the perceived frequency or familiarity that is crucial. Perhaps, given a large enough and more representative corpus, speaker perceptions and objective measure might coincide.

## Chapter 6

Chapter 6 describes a Maltese Treebank Čéplö created, adopting the Universal Dependencies (UD) Treebank annotation standard. After a brief justification on why he chose a UD-based annotation, Čéplö goes into the details of the UD annotation system and how he has applies it to the grammar of Maltese. This is followed in section 6.3 by a description of the morphological features used to annotate the data, which are relevant to the syntactic analysis. Thus, e.g., although Maltese distinguishes morphologically between singular, plural, collective and dual on certain classes of nouns, only singular versus plural are relevant in grammatical agreement, with plural and dual agreeing with plural forms and singular and collective with singular forms.

In the second part of this chapter (section 6.4), Čéplö describes his adaptation of the UD annotation scheme to Maltese, discussing in detail, and with many examples, how and why he applies the UD relation labels 'to the structure of the Maltese sentence' (p. 91). In effect, as the author himself claims, 'this amounts to compiling a rough description of Maltese syntax' (p. 91), based both on previous work on Maltese (in particular, Borg and Azzopardi-Alexander 1997 and Vanhove 1993'), and his own analyses when necessary.

This is a very commendable undertaking and the result is an insightful and well-argued description of Maltese syntax, which is itself a significant contribution made by this thesis to the study of Maltese syntax. The description is very detailed and accurate, on the whole, though, it goes without saying, that many points of grammar invite further discussion. However, this is not the place to carry out such a discussion in detail. I will only mention two points as examples.

On page 149, Čéplö says that, in the example *Xi drabi din l-istramberija tiġi assimilata* 'Sometime this strangeness is assimilated', 'an ambiguity arises between the dynamic passive and the stative passive where the place of *ġie* is taken by KIEN.' It is debatable whether the use of *ġie* or *kien* makes some 'semantic' difference (stative/dynamic), s has indeed often been claimed. To me, it appears more to be the case that *ġie* is replacing *kien* in the analytic passive, with the latter being used, when at all, in more formal registers.

Another example is on page 154, where Čéplö says that '*sejjer*...is...a little more syntactically flexible than *qiegħed*..., and unlike *qiegħed*, it can function as a proper predicate'. Indeed, as far as I can see, *sejjer* always functions as a predicate meaning 'go', and never marks the future (unlike *sa*, *ser*, which always do), which is why, e.g., it is not possible to say **sejrin imorru Għawdex instead of vs. sa mmorru Għawdex.* This is different from *qed* and *qiegħed.*

There are one or two mistranslations, as in the example on page 153, *Għax miniex ngħarfek!*, which is glossed as 'Because I don't know you!' but should be more correctly rendered as 'Because I don't recognise you!' On the whole, apart from a few odds and ends, the description of Maltese syntax presented by Čéplö is very detailed, consistent and well argued, and he deserves full credit for it. His analyses are insightful, consistent with the empirical base, and show an eye for detail and a sharp analytic mind. Finally, I believe that taking into consideration native speaker intuitions will enrich the empirical base and allow a more precise account of the data.

**Chapter 7**

In chapter 7, Čéplö provides the answers to his research questions on the basis of a quantitative analysis of the clause types described in detail in chapter 6. Apart from specific answers to the research questions, a number of interesting points which emerge from the analysis are also discussed. For example, it turns out that sentences in Maltese newspapers 'tend to be the longest', when compared to the other text types in the data-base, confirming what had been observed before in Fenech (1978). Moreover, it appears that sentences tend to be longer in written texts than in (quasi)spoken texts.

It is clear that Čéplö is very focused in his analysis and does not allow himself to be misled by the nature of the written texts. For example, on page 198, Čéplö discusses an example (4), which at first sight, because of a comma, appears to be a case of parataxis. However, he correctly concludes that 'such clauses are, after all, in no actual dependent relationship to their governor and should thus be considered syntactically equivalent to main clauses.' The fact that these appeared to be in a parataxis relations stems from the nature of the data format, i.e., the written medium, and the incorrect use of the comma (comma splice) by the writer in the text. This clearly shows how diligent Čéplö is in his analysis.

The conclusions that Čéplö reaches are consistent with his analysis of the data and well-grounded, also theoretically. Apart from *acls* (adjective clauses), Maltese clauses predominantly display SV and VO order. The main conclusion is that, within the context of the notion of discourse-configurationality, as defined by Kiss (1995), and, in particular, in comparison with Hungarian (on the basis of which Kiss developed his theory), Maltese cannot be described as a discourse-configurational, or a topic-prominent, language. Indeed, comparing Maltese to Hungarian and English as extremes on a continuum, and including Greek as an in-between, quantitatively Maltese turns out to lie closer to English, and further from Hungarian, than Greek in terms of the frequency of occurrence of SV/VO vs VS/OV. Čéplö also calculates that, assuming spoken Maltese displays a large number of VS/OV constructions, it would still not be all that much closer to Hungarian and, therefore, cannot be said to be discourse configurational in that sense, either.

The analysis is very interesting and convincing, although one would need to clarify a number of fundamental concepts to be able to make a better judgment, in particular, the concepts of topic and focus, or old and new, as well as configurationality vis-à-vis linear and hierarchical structure, and implications for constituent order. One could also question the very concept of dominant order (e.g. as opposed to basic order) and its theoretical significance. For example, the fact that the dominant order is SV and VO does not tell us what and why other orders are possible in one language but not in another, or even offer an explanation of why one order is used instead of another in a given context in a language.

Indeed, Čéplö himself shows that he is very aware of these issues when he states the fact 'that speakers of Maltese can…and do produce OV&VS sentences is a fact about Maltese', while the fact '[t]hat 'in ~75% of cases these speakers produce…SV order and in ~95% of cases they produce…VO order…is also a fact about Maltese', and '[b]oth these facts need to be reconciled in a way which not only accurately describes the data, but allows for making correct predictions, regardless of whether or not these two facts belong to different linguistic domains' (p. 241).

**Chapter 8**

The Concluding section is short and to the point.

To conclude this report, this is a very significant study which enriches both the field of linguistics, in particular, research on the syntax of constituent order, and, even more so, the relatively young field of Maltese syntax. The review of the relevant literature is broad and critical, but also balanced, concise and focused. The methodology adopted is well justified and consistently applied throughout, and the rationale for the data collection and the subsequent analyses is clearly stated. Underlying the analysis is very solid groundwork, with the discussion of the results and conclusions carefully crafted, taking into consideration potential or actual loopholes and ambiguities.

On the whole, the thesis is very well and clearly written, well structured and coherent, though occasionally, perhaps unnecessarily, slightly tongue in cheek. Formatting and referencing are consistent throughout. A relatively small number of typos and occasional inaccuracies in translation do occur but they do not, in any way, detract from the overall high level of accuracy both in content and form, and can easily be remedied. (A list of minor errors can be provided for the benefit of the candidate.)

On the basis of the evaluation above, I conclude:

1. the dissertation undoubtedly meets the standard required of a doctoral dissertation;
2. I recommend the dissertation for a public defence;
3. I propose the grade of ‘Pass’ for the work submitted.

________________

Prof. Ray Fabri
Institute of Linguistics and Language Technology
University of Malta