Posudek oponenta disertační práce Mgr. Slavomíra Čéplö

"Constituent order in Maltese: A quantitative analysis"

předkládané v roce 2018 na Ústavu srovnávací jazykovědy

I. Stručná charakteristika práce

The dissertation presents an analysis of dominant and variant orders of the main constituents (S, V, O) in various clause types in Maltese. The analysis is based on data from a dependency treebank of Maltese that has been annotated by the author of the thesis; the treebank is thus an important and useful byproduct of the presented work.

II. Stručné celkové zhodnocení práce

The dissertation is a monumental work whose impact on computational processing and corpus research of Maltese is likely to outshine its central topic and research questions, however interesting they are on their own accord. I have some reservations to certain partial decisions (and I will discuss the details below) but these cannot imperil my general impression of the work, which is definitely positive.

III. Podrobné zhodnocení práce a jejích jednotlivých aspektů

The dissertation is formally correct and well structured. The graphical layout is standard, my only complaint is that sometimes a page break separates a Maltese example from its gloss or dependency tree. Also it would have been advantageous (for those who read the PDF version of the text) if URL addresses were clickable hyperlinks.

The text is written in good English, only with a limited number of typos. It is clear and understandable, the research story is built well, starting with definition of terms, a survey of previous work, continuing with the methodology (details of data annotation) and finally presenting and analysing the results. For the most part it is clear where the author is headed and why he is doing what he is doing. A possible exception is the description of statistical analysis in R (e.g., Section 7.3.2.2.4), where it feels like a mere documentation of what the author fed the software with, without sufficiently explaining what exactly it means and why it is the right way to go.

The author works meticulously with related literature (always citing both the original and translation of non-English quotations). I cannot judge whether a piece of work relevant to Maltese syntax is missing, but the list of publications and their evaluation looks exhaustive. I found two errors though: Milička (2014) is cited twice in the text, but the corresponding full reference does not appear in Bibliography. The same holds for Shopen (2007) (on page 14(36) incorrectly cited as "Shoppen").

I find the presented research methodologically sound, except for two major deviations from the Universal Dependencies (UD) annotation standard, which also project to the observations about constituent order.

My first objection concerns the core-oblique distinction. It is a cornerstone of UD dependency classification, and it is favored over the argument-adjunct distinction, and over annotation of semantic roles. In contrast, the author lets himself get misled by semantic roles, and arrives at what is essentially the argument-adjunct distinction, although the arguments are presented as core arguments. (On the other hand, it would not be fair to solely blame the author for this

misunderstanding. The UD project itself started with a vague reference to core/oblique arguments, and the guidelines (especially v1 guidelines) were quite blurry in this respect.) Nevertheless, the consequences are not necessarily (not all of them) negative. It led the author to do extra work on valency of Maltese verbs, which will be useful in the future. I also agree that it is useful to preserve the distinction between oblique argumetns and adjuncts, despite its being considered ellusive and difficult to annotate by the UD guidelines. Fortunately, the author used a distinct label for the oblique arguments, `nmod:obj`, so it will be easy to relabel them with the optional subtype `obl:arg` defined in UD v2. Thus the only negative point is that they are misclassified as core arguments in the present work, and their statistics are mixed up with those of direct objects.

Some more detailed comments on the matter:

> Page 112 (134) and onwards; section 6.4.4.2: "Since they are semantically equivalent, should the two … be … equivalent in terms of their syntactic relationship?" The answer is definitely **NO.** This is a fundamental aspect of UD: it is not a semantic representation. The same semantic role can be expressed by a core argument (as in English *I gave **John** a book*) or by an oblique argument (as in English *I gave a book **to John***).
>
> Page 114 (136) "They [Tesnière's actants I-III] are, in all but name, what UD v1 terms core nominal dependents." Even this is not completely true because the definition heavily depends on semantic roles, something which UD strives to avoid as much as possible.
>
> Page 121 (143), the paragraph starting with "As for the second main branch…": "Moreover, such dependents typically fulfill the semantic role of a direct object (patient, VALLEX PAT)" Again, a core argument should be identified by the coding strategy and by syntactic rules that target it, such as passivization. If a reference is made to semantic roles then only to identify the primary transitive clauses in the language (see Chapter 3 by Andrews in Shopen, 2007). But here proto-patient is not merely someone who is seen or observed; it is assumed that the proto-patient is acted upon and its state is changed.
>
> Page 123 (145): "which can be blurry under the best of conditions" … this is exactly the reason why UD avoids the argument vs. adjunct distinction.

My second reservation concerns the usage of the `xcomp` relation. From the examples given, I suspect that it is not used in accordance with the UD standard, namely for clauses that do not have their own overt subject, but the subject is understood as coreferential with a core argument of a higher clause. One obvious violation of the guidelines is that xcomp clauses in the Maltese treebank often have overt subjects. I believe that in the verbal chains, the overt subject should always be attached to the highest verb in the chain, even if it occurs after the last verb. As the author notes, this decision has considerable impact on the constituent order observations in Chapter 7. (This also illustrates how crucially the word order analysis depends on the selected annotation scheme.) Furthermore, I am not convinced that in some cases the subject is really obligatorily inherited from the higher clause, and not just incidentally coreferential with the higher clause in the particular sentence (see also my questions below).

The main contributions of the dissertation are two: (1) this is the first assessment of constituent order in Maltese based on real corpus data (2K sentences / 44K tokens from various text genres and domains). The results differ from the claims published in previous work, and the quantitative grounding makes the new results more trustworthy. (2) As a byproduct, a large tagged corpus of Maltese texts, and a small (but the first available) dependency treebank of Maltese was created.

Both were prepared by the author during the work on the thesis (documentation of annotation decisions is part of the dissertation).

IV. Dotazy k obhajobě

Selected questions for the defense:

- The author explicitly excludes translations into Maltese from his corpus, yet he includes proceedings of the Parliament of Malta. Is the Maltese Parliament really monolingual, i.e., no translations from English?
- Page 78 (100), section 5.4.1.3.34 QUAN is empty. Why? And what should be here?
- Page 134 (156), example (66): Why is there an `xcomp` between *kburin* and *Laburisti*? Is it ungrammatical to say something like *aħna kburin li huma Laburisti* "we are proud that **they** are Laborists"?
- Page 122 (144), example (43): This is an interesting example that would deserve more discussion. The patient *ritratti* is coded similarly to core arguments (bare noun phrase) so it would almost deserve to be labeled as `dobj`, but it would obviously be strange in a passive clause. On the other hand, one should rule out the possibility that it is subject, using some tests of subjecthood that are applicable to Maltese.
- Page 124 (146): Subject marked by the object marker *li*. What makes it a subject then (except its semantic role, which is irrelevant in UD)?

V. Additional remarks

The following comments are not so important to be necessarily discussed during the defense, yet they might be useful for the follow-up work.

- Page 1 (23). The aim is to be descriptive, theory-neutral. In absolute terms, this is hardly possible because even UD has some theoretical assumptions (despite claiming that it is not a linguistic theory), and if not UD, then definitely VALLEX, with its FGP-based roots.
- Page 63 (85), section 5.3.3.2. I don't like the decision to normalize all quotation marks to U+0022, as it throws away the information on directionality (opening vs. closing mark). Same for hyphens vs. longer dashes.
- Page 83 (105): There is some confusion about the statistics of individual UD releases. UD 2.1 has 60 languages *without* Maltese (which was not released in UD 2.1). Number of "planned languages" is unimportant, as it changes frequently, and some languages have been "planned" for years, without any data arriving. The UD 2.2 release is scheduled for June 2018 and will have at least 69 languages (without Maltese). The November 2018 release will be UD 2.3.
- Page 85 (107): The number of UPOS tags is 17, not 16; CCONJ is in UD v2 (in v1 it was CONJ). I understand why the dissertation prefers to work with the Maltese-specific POS tags, but in the CoNLL-U format these should be in the XPOS column, and the UPOS column should hold the coarse-grained universal tags instead.
- Page 94 (116), table 6.5; and then page 154 (176), section 6.4.4.8.6: the `part` relation cannot be used in UD in this form. It must be a subtype of a defined universal relation, perhaps `aux:part`?
- Page 95 (117), example (1): *tagħkom* should be attached as `nmod:poss`, not as `case`.
- Page 96 (118): Why is *mill-* attached to *Ministru* and not to *Mizzi*? (Similar example (6) is annotated correctly.)
- Page 99 (121), rule c: "but only if the valency of the verb … allows an agent noun phrase" – how exactly is agent defined for the purpose of this thesis? Are there any verbs other than those describing weather conditions that do not license an agent?

- Page 107 (129): The author rejects the analysis of subordinate clauses as `csubj` in copular sentences. I believe that the `csubj` analysis in (26) is quite correct and the fact that the subordinate clause in (27) is `advcl` does not contradict it.
- Page 131 (153): "if *baqa'* were treated as an auxiliary…" – I don't see *baqa'*.
- Page 132 (154): "nsubj … attach to the first or last verb in the chain" – should it be "the highest verb"?
- Page 133 (155), example (64): I do not understand why *jkun* is not a copula.
- Page 135 (157), example (67): Why isn't the clause csubj, instead of ccomp? (And if it is, then it influences the constituent order, although the chapter 7 in this work only looks at nominal subjects.)
- Page 137 (159): examples (70) and (71) seem OK to me without `dislocated`. Example (72) looks like a good case for `dislocated`.
- Page 141 (163), example (78). Why does not the quantifier *kemm* (how much) depend on *drawwiet*?
- Page 152 (174), example (96): The negative pronoun should be `nsubj`, not `neg`.
- Page 161 (183): Why is the number *wieħed* "one" attached as `det` and not `nummod`?
- Page 203 (225), figure 7.8: The caption does not explain what is 0% and what is 100%.
- Page 203 (225), 7.3.2.3: The statistics of dobj and nmod:obj should not be mixed together. (Especially if iobj is singled out.)
- Page 222 (244): Example (11): This should be `ccomp`, not `xcomp`. Example (12): This probably is `xcomp`, but the subject should be attached to *jista'*.

VI. Závěr

The presented work meets the requirements of a dissertation, therefore I recommend it to the defense and my tentative assessment is *passed*.

Předložená disertační práce splňuje požadavky kladené na disertační práci, a proto ji doporučuji k obhajobě a v předběžně ji klasifikuji jako *prospěl*.

22.5.2018 Daniel Zeman