Problem-oriented Corpus Annotation and the Hebrew Bible

Johan de Joode
KU Leuven, Leuven, Sint-Michielsstraat 4 - bus 3100, Belgium
johan.dejoode@kuleuven.be

Abstract: In this contribution, I argue that the exegetical and stylistic study of the Hebrew Bible would benefit from the creation and storage of qualitative and quantitative annotations using problem-oriented corpus annotation (de Haan 1984). Within Biblical studies exegetes are used to static interfaces which allow them retrieve information, but not enhance it with anything more elaborate than user notes. I present a roadmap for the development of an annotation tool tailored to the Hebrew Bible with the sole objective of enriching the data that is already present in open source datasets like that of the ETCBC. Based on my experience with a dataset to annotate conceptual metaphors in the book of Job (bibliametaphorica.com) and the literature on corpus annotation (Sinclair 2004; Leech 2005; Fort et al. 2012), I argue that data creation and enrichment is a challenging, yet rewarding endeavour. It is challenging because it is circular, viz. labels are informed by the data are hence difficult to a priori define. Furthermore, it is difficult to be consistent and the actual, manual labelling of the text requires interpretative choices that cause editorial fatigue. Fort et al. (2012) suggest that annotation campaigns have differing degrees of difficulty which can be mitigated not just by inter-annotator rating, but by conscious decisions to lower the annotation complexity. A user interface is needed that limits annotation complexity and that allows researchers to annotate the text with minimal effort. The end-result is an XML document, for instance, that contains both the text and the annotations, in a format that can be merged back into the original database, but need not be. The existence of a tool for the manual annotation of open data will increase the replicability of research as well as its democratisation, as students worldwide can create and share their data.

Keywords: corpus annotation, corpus linguistics, quality assurance, Hebrew

Introduction

Since the advent of computer-based studies of the Hebrew Bible, the semi-automatic annotation of the digitised text has played a significant role: the basic text of the Bible was annotated with a variety of scholarly observations including, for instance, a critical apparatus, lexemes, morphological and syntactic analysis, discourse properties, and word-sense disambiguation. These annotations have enabled scholars to retrieve information rapidly and often exhaustively; they have opened avenues for quantitative and qualitative inquiries, and, equally importantly, they have raised new questions. The commercial potential of these annotations, especially for students of the original languages, has broadly been explored, but generally tends to take the text (and its readers) hostage.[1] The data is safely stored, yet only accessible to the fortunate few. I believe that new research on the Hebrew Bible is dependent on encouraging researchers and students alike to contribute to the data creation process through corpus annotation.

[1] There is to date no single, exhaustive dataset for the Hebrew Bible that can be used for commercial and non-commercial applications alike without any restriction (MIT, or BSD licensed, for instance). The commercial value of the annotations has outweighed the scholarly liberty to share the data, and the people’s right to access it. One can assume that the commercialisation of the data is a direct consequence of the extensive efforts required to annotate a text.

The annotation of a text requires the preparation of the text under scrutiny, the definition of an annotation scheme, the development of tools to annotate the text, the definition of extensive quality controls, the documentation of the features, and the distribution of the data. This is a complex, arduous, and time-consuming enterprise which often leads to idiosyncratic databases that fail to provide accessible, consistent, usable data in a variety of shapes and formats. These idiosyncrasies limit scholarly research. There are few, if any, comparisons and evaluations of the quality of the data across datasets. Indeed, each dataset is so idiosyncratic that comparisons are difficult, if possible at all. In line with this volume's aim to reflect on the creation of original language resources for translators, students, and scholars alike, this position paper advocates for the strategic development of a set of tools that enable the enhancement of the current data and the creation of new data through corpus annotation. The main point is that research will benefit from the conscious attention for the quality of the data as well as its interchangeability.

The Challenges of Corpus annotation

Leech (2004) defines corpus annotation as "the practice of adding interpretative linguistic information to a corpus."[2] He cites part-of-speech tagging as an example, viz. a task that is typically performed on all of a text's tokens. There is a different type of annotation, however—the one that is my focus—namely problem-oriented annotation (de Haan 1984). The latter is "the phenomenon whereby users will take a corpus, either already annotated, or unannotated, and add to it their own form of annotation, oriented particularly towards their own research goal" (McEnery and Wilson).[3] Whereas Sinclair (2004:190-191), one of the pioneers of corpus linguistics, warned against approaching a text through its annotations, it is now widely accepted to use problem-oriented annotation for the study of specific research questions and the enrichment of the data, for instance for stylistics (McIntyre 2007:572-574). Within Biblical studies, however, problem-oriented annotation is a rather marginal phenomenon. In what follows, I describe three ways of overcoming these lacunae. I also describe the difficulties inherent to annotating corpora and I advocate for best practices.

Corpus annotation is a standard matter in introductory courses of corpus linguistics, but in practice it is a challenging enterprise, not at the least because it requires sequential interpretative decisions. Fort, Nazarenko, and Rosset (2012) measure the difficulty of annotation campaigns using the following six dimensions: the discrimination and delimitation of the units to annotate, the expressiveness of the annotation language, the tagset dimension, the degree of ambiguity, and the context. The first two dimensions describe what to annotate, the next three how to annotate it, and the last dimension covers both. Space does not permit a detailed discussion of all these dimensions, but their insights can be practically applied in order to reduce the difficulty of the annotation, and thereby increase its accuracy.

Problem-oriented annotation can be made more straightforward if the annotator does not have to discriminate or delimitate the units of analysis. This is not possible for all types of analysis, but wherever possible the units of analysis should be predefined: "the simplest solution is to rely on an obvious segmentation which can be automatically computed" (Fort, Nazarenko, and Rosset 2012:900). In practice it is often possible to start from a query. In cases like semantic role-labelling, however, this discrimination is challenging.

[2] http://ota.ox.ac.uk/documents/creating/dlc/chapter2.htm

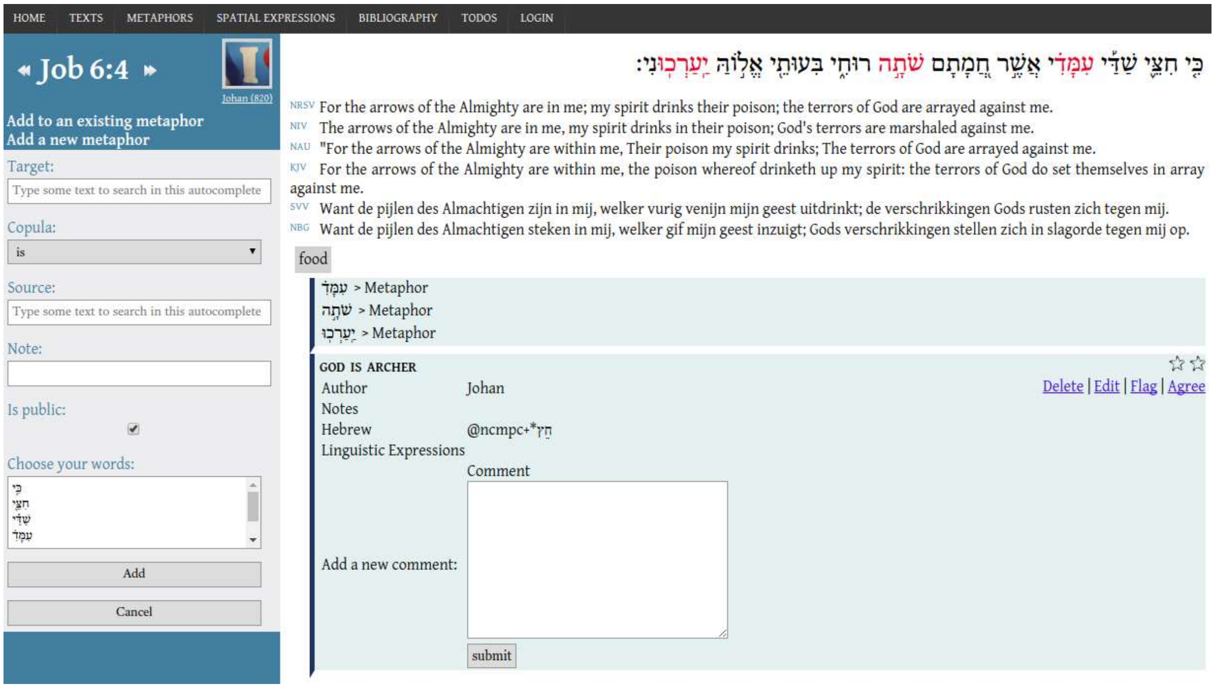
[3] http://www.lancaster.ac.uk/fss/courses/ling/corpus/Corpus2/2PROBLEM.HTM This website provides additional information to McEnery and Wilson (2001).

The recommendation to start from a query might seem obvious, but annotators new to the task could be tempted to annotate a text, rather than a set of predefined features. The original prototype I developed for Biblia Metaphorica, intended to provide a system to annotate conceptual metaphors in the Hebrew Bible (Figure 1). It did not include a query engine, and hence it made the annotation task more complex as the annotator had to segment the text during the annotation. The preoccupation with segmenting is shared by the Pragglejaz group (2007) which, in its methodology to tag metaphors, starts with segmentation or unit delimitation as the first analytic step. In hindsight, problem-oriented tagging is ideally done in multiple 'passes' over the data: rather than annotating a text, one can annotate all nouns, subsequently all verbs, or other query results relevant to the research question under scrutiny. If one wanted to label metaphors in a text, one could start by annotating all nouns, for instance, with labels such as 'animate/inanimate', which would reveal cases where inanimate nouns accompany verbs that traditionally have animate subjects (for instance, 'Hope left the building').

Fort, Nazarenko, and Rosset argue that the annotation should partitioned in elementary annotation tasks which should be a simple as possible. This has the additional benefit that certain parts of speech occur in similar contexts, thus making interpretative decision easier. Also, one can hence track the progress of the annotator easily. It also encourages the annotator to postpone large-scale interpretative observations. It forces the analyst to distance herself/himself from the data. Finally, it also highlights issues with the annotation scheme sooner rather than later.

Just as the delimitation and discrimination of the textual features to be annotated should be limited, so too the variety of choices within an annotation scheme should minimal; the options could even be binary. Bayerl and Paul found that inner-annotator agreement is correlated to the "number of categories in a coding scheme" (2011:699). Fort, Nazarenko, and Rosset (2012:900) observe the effect of the tagset dimension, and the degree of ambiguity of the different tags. Indeed, problem-oriented annotations are not notes, they are data. Many of our current tools, including Shebanq and BibleWorks, to name but a few, allow researchers to add notes to the text using text fields. It is recommended, however, to predefine an annotation scheme that has fixed labels. This yields consistent results and avoids editorial fatigue when annotating. If needed, one can use a training corpus different from the corpus one wants to analyse to define and document the annotation scheme. Again, this can be an incremental method that has multiple iterations.

Once an annotation scheme is defined, it should be enforced. For instance, if one wants to annotate corporeal metaphors in a text, one could first label all corporeal language, subsequently whether it is literal or metaphorical. By limiting the options, the task is simplified for the annotator. Also, the advantage of binary tagsets is that they can be useful for questions one does not have yet at the time of annotating. A later analysis could evaluate whether these corporeal metaphors are deliberately used in the literary context, for instance. This does not only hold for metaphor analysis, the same hold for an analysis of, say, the emotional affect or valence of adjectives in the Psalms. Along these lines, one can use a single label for features that need to discussed or revisited, viz. "needs more reflection." Using a limited tagset has the advantage that it forces the annotator to predefine the labels; this speeds up the process considerably, especially when compared to the type of annotation that requires the annotator to type actual labels, or worse, invent them on the spot. The overall intention is to simplify the process by reducing the interpretative decisions of the annotator. The development of a tagset that is carefully crafted and well-documented, with no ambiguous labels, is of great value.



Target: Type some text to search in this autocomplete

Copula: is

Source: Type some text to search in this autocomplete

Note:

Is public: ☑

Choose your words:
כִּי
חִצֵּי
שַׁדַּי
עִמָּדִי

כִּי חִצֵּי שַׁדַּי עִמָּדִי אֲשֶׁר חֲמָתָם שֹׁתָה רוּחִי בִּעוּתֵי אֱלוֹהַּ יַעַרְכוּנִי׃

NRSV For the arrows of the Almighty are in me; my spirit drinks their poison; the terrors of God are arrayed against me.
NIV The arrows of the Almighty are in me, my spirit drinks in their poison; God's terrors are marshaled against me.
NAU "For the arrows of the Almighty are within me, Their poison my spirit drinks; The terrors of God are arrayed against me.
KJV For the arrows of the Almighty are within me, the poison whereof drinketh up my spirit: the terrors of God do set themselves in array against me.
SVV Want de pijlen des Almachtigen zijn in mij, welker vurig venijn mijn geest uitdrinkt; de verschrikkingen Gods rusten zich tegen mij.
NBG Want de pijlen des Almachtigen steken in mij, welker gif mijn geest inzuigt; Gods verschrikkingen stellen zich in slagorde tegen mij op.

food

עִמָּדִ > Metaphor
שֹׁתָה > Metaphor
יַעַרְכוּ > Metaphor

GOD IS ARCHER
Author Johan
Notes
Hebrew @ncmpc+*חֵץ
Linguistic Expressions
Comment
Add a new comment:

Figure 1. Example of the original prototype for www.bibliametaphorica.com

Strategic Annotation

The prototype of *Biblia Metaphorica* (Figure 1) did not benefit from the above observations, rather on the contrary: although annotators could reuse labels, they were not defined in advance, and the unit of analysis was the text, rather than specific segments. Based on my experience with *Biblia Metaphorica* and the development of the annotation module for the *CLiC Dickens* project (Mahlberg et al. 2016), this section highlights three pathways to promote problem-oriented tagging: modularity, simplicity in user interface design, and quality control.

Leech defines several standards for corpus annotation, the first of which is "annotations should be separable." The separation of text and annotation is less important now compared to several decades ago, as one can easily strip text from its annotations (McEnery and Hardie 2012:13). In our case, however, the separation of annotation and text is valuable as we are all working on the same corpus, rather than on a variety of different corpora. The peculiarity of our research is that it focuses on a small corpus. The Hebrew Bible only contains ca. 440k tokens. Hence, I suggest that annotation should be modular: we should separate text and annotation, not because we would not be able to remove the annotation of the text, or only extract the annotations, but because we should treat annotations as annotations and not as linguistic data. Modularity is the idea that the base text with book, chapter, and verse divisions (or for the scrolls: scroll, column, and line numbers) is centrally distributed with unique identifiers for each token (where a token can be a character or a 'word'). The identifiers can be used to attach annotations to the text. The main text should be in human readable format, such as XML, thus also allowing for better version control (compared to version controlling binary files). If annotations are modular, and the text is distributed separately, it will be possible to merge multiple annotation sets by their shared reference to the same object identifiers.[4] This requires that new iterations of, for instance, the ETCBC database use a system that connects object identifiers to their earlier identifiers if they change over time. Although the separation of text and annotations is not a conditio sine qua non for problem-oriented tagging, it is first step towards the interchangeability of data.

The second improvement to current tools is the user interface. As highlighted above, the annotation process should be as easy as possible, distinguishing multiple passes over the same data whilst avoiding the manual creation of new labels during the annotation. The user interface should therefore focus on the two stages of the annotation: the definition of the annotation scheme and the actual annotation. The first stage requires an interface that can handle practical guidelines for annotators. In Leech's (2005) words, "detailed and explicit documentation should be provided." The documentation requires both straightforward and complex examples. It is not an afterthought, rather it should just be a click a way during the annotation. In order to facilitate this type of research, the annotation itself requires a user interface that does not require shifting between the mouse and the keyboard, for instance. More often than not, there are many cases that are easy to annotate, and these cases should be rapidly dealt with. The actual annotation interface should show the progress of the annotation and should be tailored at the specific annotation scheme (this can be done automatically). It should hence not come as a surprise that the user interface can encourage best practice or distract from them. Bayerl and Paul (2011:699) find that the accuracy of manual annotations depends, among other factors, on whether the annotators have received training and the intensity of that training. The user interface can provide prefatory tests to evaluate the annotators' mastery of the annotation scheme.

The third and final improvement to current linguistic annotations would be the use of measures for quality assurance to evaluate the exhaustiveness, quality, and inner-annotator agreement of an annotation campaign. The original prototype for *Biblia Metaphorica* contained primitive buttons to express degrees of certainty for an annotation and agreement with an annotation provided by another annotator. Such measures are too simplistic and they should be avoided. Quality can be controlled using quantitative measures such as *Kappa* (Carletta 1996) or more simplistically the proportion of agreement. We could also apply these measures to existing morphological, syntactic, semantic, and discursive annotation in the ETCBC database. If only a single set of annotations is available, the inter-annotator agreement cannot be computed, but the consistency of the annotation (for instance for identical lexemes in different context can be computed). For automatic annotations, which are not the primary focus in this contribution, one can manually compile a gold standard, e.g. a dataset with annotations that are in perfect agreement with the annotation scheme, but that are tagged manually. That gold standard can be used to compute precision and recall, or a balanced *F measure*, for instance (Manning, Ragahvan, and Schutze 2009:156).[5] The results of both the gold standard and the annotation can be distributed in a format that is both readable by humans and computers.

Conclusion

This contribution has argued for the creation of tools for problem-oriented corpus annotation. It has provided procedures required for such annotation as well as suggestions for the creation of versatile tools. This position paper aims to be a theoretical contribution to the debate in order to aid the development of a minimal-viable product for problem-oriented annotation. Annotating literary texts is a

[4] On the identifiers, cf. Tauber's suggestions to extend Strong's numbers at https://www.academia.edu/35220175/Linking_Lexical_Resources_for_Biblical_Greek.

[5] This does leave us to wonder whether the copyright of the data resides.

challenging enterprise. It is not a glamorous task and it is often glossed over to get to the 'real data.' What can be achieved with manual annotation is, nonetheless, valuable: annotation is essential for new corpus linguistic studies and it can set the standard for peer review as researchers can be asked to publish their data. Problem-oriented annotation is valuable because it breaks down interpretative processes into smaller steps that can be verified and shared. This will affect the study of style, for instance, with its quantification of stylistic features and its desire to discover patterns, but it can also be used to generate corpus-based ontologies, frame semantic analyses, or metaphor studies. The tools developed for problem-oriented corpus annotation are also likely to become significant when fine-tuning our automatic annotation methods for use by machine learning techniques that require labels. Problem-oriented annotation need not necessarily take years: a carefully crafted annotation campaign can take place rapidly and it can even be tailored to novice annotators.[6]

This contribution has proposed several protocols for the creation of new data, and the evaluation of data generated by others. In summary, (1) one needs to fix an annotation scheme in advance, if necessary multiple, simple iterations are preferred over a single, more complex iteration. (2) One should segment the text by selecting features and delimiting them; (3) interpretative decisions regarding the text should be delayed as one separates the analysis of the annotation into a data collection and an interpretative stage; (4) the process should be made as easy as possible: ideally annotation takes place with a single hand, and with tools that help to define and document the annotation scheme with specific examples. The development of a tool should not primarily focus on tasks that can be done with existing tools or custom scripts. An annotation campaign requires a planning phase to determine how to segment the text, what definitions are used, what the value of the data is, and how quality will be assured. It is believed that careful annotation campaigns will provide for new research questions and advanced insights into long-standing issues.

References

Bayerl, Petra Saskia, and Karsten Ingmar Paul. (2011) "What Determines Inter-Coder Agreement in Manual Annotations? A Meta-Analytic Investigation" *Computational Linguistics* 37:4, 699–725. Available online from http://www.mitpressjournals.org/doi/10.1162/COLI_a_00074

Carletta, J. (1996) Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics* 22:2. Available online from http://homepages.inf.ed.ac.uk/jeanc/squib.pdf

de Haan P. (1984) "Problem-oriented tagging of English corpus data." In *Corpus linguistics,* J. Aarts and W. Meijs (eds.). Amsterdam: Rodopi.

Fort, Karën, Adeline Nazarenko, and Sophie Rosset. (2012) "Modeling the Complexity of Manual Annotation Tasks: A Grid of Analysis." *Proc. COLING 2012 Tech. Pap. Pages* 895–910, COLING 2012, Mumbai, December 2012. December (2012): 895–910.

Leech, G. (2005) "Adding Linguistic Annotation" in *Developing Linguistic Corpora: a Guide to Good Practice.* M. Wynne (ed.). Oxford: Oxbow Books: 17-29. Available online from http://ota.ox.ac.uk/documents/creating/dlc/

Mahlberg M., Stockwell P., de Joode J., Smith C., O'Donnell M. (2016). "CLiC Dickens: Novel Uses of Concordances for the Integration of Corpus Stylistics and Cognitive Poetics." *Corpora, 11* (3), 433-463.

Manning, Christopher D., Prabhakar Ragahvan, and Hinrich Schutze. (2009) "An Introduction to Information Retrieval." *Information Retrieval.* Cambridge: Cambridge University Press. Available online from http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf

[6] Amazon's *Mechanical Turk* can make this even more scalable.

McEnery, Tony, and Andrew. Hardie. (2012) *Corpus Linguistics Method, Theory and Practice.* Cambridge: Cambridge University Press.

McEnery, Tony, and Andrew Wilson. (2011) *Corpus Linguistics: An Introduction.* 2nd edition. Edinburgh: Edinburgh University Press.

McIntyre, Dan. (2007) "Trusting the Text: Corpus Linguistics and Stylistics." *International Journal of Corpus Linguistics* 12:4, 563–75.

Pragglejaz Group. (2007). "MIP: A Method for Identifying Metaphorically Used Words in Discourse." *Metaphor and Symbol* 22:1, 1-39.

Sinclair, John. (2004) *Trust the Text: Language, Corpus and Discourse.* London: Routledge, 2004.