Papers for the Copenhagen Workshop on Open Biblical Resources

Nicolai Winther-Nielsen
Fjellhaug International University College Denmark, Leifsgade 33.6, 2300 København S
and Eep Talstra Center for Bible and Computer, VU University Amsterdam
nwn@dbi.edu

Abstract: A workshop on open resources for the original languages of the Bible in Copenhagen in March 2018 was the start of a new Copenhagen Alliance for Open Biblical Resources. The point of departure for the workshop was the need for programs and applications like Paratext and Bible Online Learner to have access to high-quality and reliable open data in order to assist Bible translators, teachers and students of Biblical Hebrew and New Testament Greek. The publication of contributions presents papers on methods for annotation, resources tracing patristic quotations and data for detached constructions in Biblical Hebrew. Reports cover tasks and data for Bible translation and research, treebanks, and applications like STEPBible and Bible Online Learner.

Keywords: Bible translation, open resources, technology, education

This volume of HIPHIL Novum publishes updated versions of the proceedings of the workshop that launched the Copenhagen Alliance for Open Biblical Resources (http://copenhagen-alliance.org). With funding from Every Tribe Every Nation (https://eten.org/), the workshop was organized by Global ICAP Coordinator, Dr Reinier de Blois, from the United Bible Societies and the Vrije Universiteit in Amsterdam, and HIPHIL Novum editor, Professor Nicolai Winther-Nielsen, from Global Learning Initiative. The former has been responsible for developing the PC program Paratext which is the main program used by virtually all Bible Translators in the world, while the latter coordinates the development of the Bible Online Learner application, designing persuasive Biblical Hebrew language learning.

The goal set for the 23 prominent international experts and young researchers convening in Copenhagen March 19-23 was to discuss and share work on original-language, digital resources that are created and made freely available for Bible translation and education in Biblical languages. All the presentations were made available as video-recordings, but participants were also invited to submit either scientific papers for peer review or short reports on crucial projects. The overall goal was to help programmers and researchers in translation and teaching technology to collaborate on creating open digital resources for Hebrew, Aramaic, and Greek. Such data are crucial for the operations of United Bible Societies, Summer Institute of Linguistics, Unfolding Word and hundreds of organizations working within the framework of Wycliffe Global Alliance or national Bible translation agencies, as well as for the Global Learning Initiative. Digital data and technologies are becoming increasingly important for global mission in a vision to provide better Bible translation and enhance theological education for the global church.

The most significant outcome of the discussions at the workshop was to form the Copenhagen Alliance for Open Biblical Language Resources as "a diverse coalition of organizations, institutions, and individuals with a common interest in making biblical language data free and openly accessible to anyone for research, language learning, translation, and other uses." This important initiative will hopefully bring many projects with common interests together and help speed up the huge task of

translating the Bible into thousands of local languages. We expect that this initiative will provide better data to be shared among technologies for translation of the Bible and for education of Bible translators and students of Theology.

Another outcome is the publication of contributions in this special volume. Two academic papers were submitted for peer review, and a third peer review paper was added, and in addition ten reports present projects and aspects of projects. Some of these contributions are published as submitted for publication in the Spring of 2018, while others have been updated before publication in 2019. The two peer-reviewed papers deal with the construction and testing of research technology, while the reports introduce prominent projects and discuss needs and goals facing us in technology-enhanced translation and education. The third peer reviewed paper focuses on the problem of relatively free phrase order in Biblical Hebrew.

A number of papers focus on the Hebrew Bible and use the *Biblica Hebraica Stuttgartensia Amstelodamsis* text, or in short BHSA, which is by now the accepted name for the database of the Hebrew Bible maintained by the Eep Talstra Center for Bible and Computer (ETCBC). This database is at the core of many projects and is currently the only open data for the Hebrew Bible. This database is also used in the two applications that were at the forefront in the invitation to the workshop, Paratext and Bible OL. In this volume Paratext is not described in a publication, but the last three papers in the volume are directly related to the Bible Online Learner application. The majority of papers are directly related to data and technologies that are crucial for linguistic research and translation, while the latter three have a clear educational focus.

The first peer-reviewed paper is by Dr Johan de Joode of KA Leuven University who did his dissertation on metaphorical language in the book of Job. In the paper submitted for the conference, "Problem-oriented Corpus Annotation and the Hebrew Bible", he presents a research project to design stylometric analyses for Classical Hebrew. He argues that exegetical and stylistic study of the Hebrew Bible could benefit from qualitative and quantitative annotations that are provided through problem-oriented tagging. Presenting a roadmap for the development of an annotation tool that does this, he wants to enrich open source datasets in a consistent and effective way. Through this interface the researcher can create an XML document with the text and its annotations, and this new data can be imported back into the original database or serve as an alternative version.

The second peer-reviewed paper, "A Critical Examination of the Intertextual Phrase Matching Module in the *Thesaurus Linguae Graecae* and Its Relevance for Biblical and Patristic Studies" is by Ernst Boogert, PhD Candidate of New Testament Studies at the Protestant Theological University in Amsterdam. He focuses on how to validate the results produced by a program to extract quotations in ancient Greek sources in the *Thesaurus Linguae Graecae* (TLG). Examining the tracking of quotation in the n-grams module (Intertextual Phrase Matching) and its algorithms, he scrutinizes the results obtained for John in relation to the *Paedagogus* of Clement of Alexandria, Matthew, and the complete works of Plutarch. He concludes that the modules deal very well with longer quotations with no or very few interpolations, but short quotations are missed, and adapted quotations are handled poorly. The suggested solution is to ignore stopwords and to allow for foreign words in its n-grams, as well as to allow for manual improvement of the algorithm's mechanisms and parameters.

The third peer-reviewed paper, "Is Biblical Hebrew a non-configurational language? Reconsidering the evidence from discontinuous phrases", is by PhD-student Marianne Kaajan who is a member of the ETCBC-team at the Vrije Univesiteit in Amsterdam. Her contribution illustrates how the BHSA text can be used for sophisticated linguistic research. The paper addresses the problem whether discontinuous phrases in Biblical Hebrew are evidence of a non-configurational language type with free word order in clauses and phrases. Kaajan reviews the literature and the evidence and proposes

that it is more likely that Biblical Hebrew has constructions with afterthought (or right dislocation) and extraposition that are not as such evidence of non-configurationality.

The report section opens with two reports on the need for open data in Bible translation and the challenges in the field. The first paper probes into the requirements that must be met in the field of Bible translation. Associate International Translation Coordinator of SIL International, Paul A. O'Rear, brings experience both from Bible translation and from the software industry into his current position. In his paper "Towards a Common Aim and Framework for Tools and Research in Support of Bible Translation and Biblical Language Online Learning", he describes the current needs of Bible translators and pastors in the Majority world context. These needs are not fully met by individual and uncoordinated research and projects presented in the Global Education and Research Technologies (GERT) section of the Society of Biblical Literature (SBL) from 2013 to 2019, nor by the activities and projects presented on the Biblical Humanities platforms. In order to set a new agenda, he demonstrates through Bible translation resources from SIL how Bible scholars could contribute to Bible translation, pointing out that synthesizing the specialized research and data of the scholarly community is the most crucial condition for progress. SIL is now increasingly focusing on training mother-tongue translators.

The next paper, "Adapting Digital Biblical Resources for Translation and Teaching: Harvesting low-hanging fruit and growing next generation resources" by Jonathan Robie, Program Manager, Paratext Ecosystem at SIL International, follows up on this practical call for data. Coming from the business world and creating XML code and query specifications, Robie has worked for many years to find solutions that meet the needs of the translators. He was the driving force in setting up biblicalhumanities.org in 2013 in order to build a community to support the use of free high-quality resources by software developers. While many resources exist, community building around them is not enough to get the data widely used. The new development is to use Jupyter Notebooks and expose them at academic conferences, but the next step must be to focus on co-participatory design. The experience and needs of users should inform the creation of the next generation of resources for translation, language learning, and scripture engagement.

The remaining reports introduced projects and research. A paper by researchers and translation specialists, Jesse A. Griffin and Todd L. Price, "Unrestricted Original Language Resources", exemplifies how the use of resources is handled in one particular project. In their unfoldingWord project, they approach Bible translation through some 50 or so Gateway Languages. Mother-tongue translation can work from these intermediate languages if these speakers have access to original language resources in these languages and have the legal freedom to use, translate, and modify them to suit their context. They suggest that Creative Commons Attribution will work for their model of a Church-Centric Bible Translation (CCBT). They illustrate their work through their new tool for checking Bible translations, translationCore, which has access to content of specific open source nature for the original texts, lexica and reference grammars.

Two projects are currently developing technologies that have huge potential for machine-translation of the Bible in the future. Global Bible Initiative (GBI) began in 2004 as a Bible translation agency producing the Chinese Standard Bible. In the process of developing technology for translation of this version, they have built Hebrew Old Testament treebanks and Greek New Testament syntactic treebanks through parsing and manual check by scholars. Andi Wu, VP of Linguistic Research at GBI in his report "GBI Treebanks as a Resource for New Applications" describes how these trees are phrase constituents that GBI uses for interlinear display of the original languages, to create concordances, and to offer translation memories on all hierarchical language levels of Hebrew and Greek. In current

projects the GBI use trees to improve the accuracy of automatic word alignment and to develop interactive machine translation of the Bible. They predict that interactive machine translation will speed up Bible translation without compromising quality.

The STEP Bible project aims to offer open data on the Bible for the global church. Senior researcher PhD David Instone-Brewer's presentation "Tyndale STEPBible Data development for machine analysis and computational linguistics" gives a useful overview of current cutting-edge achievements of their work. The development of STEPBible by Tyndale House Cambridge has created datasets which are directly useful for computational linguistics. Moreover, some of these datasets have interesting potential for machine translation due to the more stringent and varied ways in which data has been refined and presented, and one project has focused on automatically tagging Bibles to Greek and Hebrew.

The last reports focus on how the BHSA is used for data creation and education. Dirk Roorda, researcher at Data Archiving and Networked Services in Den Hag, explains how new data can be created in his paper "Text-fabric: handling Biblical data with IKEA logistics." The Text-Fabric data set format is a minimalistic model to represent text in a way that allows for adding arbitrary information at all textual levels, precisely and firmly anchored. Using an IKEA ware house as an illustration, he recommends reuse and repurposing of data, and then reimport to the original data with a minimum of interference or duplication of work. In this way Text-Fabric can do data construction work, as the website SHEBANQ indicates. In 2018 he has focused on creating treebanks from the BHSA data and comparing them with the morphology of other data sets.

The use of the BHSA in the paper by PhD candidate at the Vrije Universiteit Christian Canu Højgaard, "Semantic mapping of participants in legal discourse", illustrates how the proposal by Robie and Roorda to use Text-Fabric and Jupyter Notebook works in participant research on the Hebrew Bible. Exploring semantic roles and participants involved in the legal material of Leviticus 17-26, he demonstrates his results for Lev 25:23-28. In three successive steps he creates a semantic map of the participants and their internal relationships: a Role and Reference Grammar account is followed by participant tracking data emerging out of Eep Talstra's pioneering research, and this is mapped semantically into participant relations.

The use of the BHSA for linguistics and for language learning is reported by Professor of the Hebrew Bible and Informations- and Communications Technology, Nicolai Winther-Nielsen, an affiliated of the ETCBC. In the paper "Interfacing the Hebrew Bible: past, present and future applications for the BHSA", he describes how the Role Lexical Module was designed as an interface in 2004-2009 to help linguists inspect the data of the Hebrew Bible in the format used in Role and Reference Grammar. Even if discontinued, the project was able to create new data on the logical structure of predicates and to test parsing rules. The next case is the learning technology Bible Online Learner which has been developed for persuasive corpus-driven language learning. environment Bible Online Learner offering a persuasive interface for enquiry and practice in and with the BHSA text. Especially the later project has provided ample evidence on how the development of user-friendly interfaces is an important aspect of how to better use open resources.

The last two papers derives from the development and implementation of Bible Online Learner. The BHSA can be linked to external resources, scaffolding the Bible with existing data and for the future. Claus Tøndering in "Adding Links to the Bible" explains how cross-references in the Bible through *hyperlinks* can link to virtually any resource. Bible Online Learner (Bible OL) currently links to pictures in the *EuroPLOT Resources* database and a Bible verse or a lexeme can link to any URL. Developing anchors for future electronic Bible study tools will allow us to reach beyond traditional Bible references and to use proper names, genre-specific text information and many other new kinds of

knowledge as links. Manual entry can be expanded into one-time harvesting, or even continuous harvesting.

Last, but not least, the BHSA can be delivered to the Majority World for training of translators and students of Theology. PhD candidate at Aalborg University, Johannes Gottschalk, in his paper "Sustainable e-learning for theologians in hard-to-reach areas: The Bible Online Learner App" suggests a solution for the challenge of the internet in countries like Madagascar. It is possible to create a sustainable e-learning solution by using Raspberry Pi mini-computers as servers and to develop a native Android App in order to provide Bible Online Learner Network Kits. This solution offers a mobile access point without an internet connection. For the client-side development he pays special attention to the challenges of computer-illiterate students by offering picture-based logins. Power banks provide uninterruptable power supply and a solar power kit will enable e-learning in hard-to-reach areas in Africa.