Proceedings of the Second Workshop on

Corpus-Based Research in the Humanities

CRH-2

25-26 January 2018 Vienna, Austria

Editors:
Andrew U. Frank
Christine Ivanovic
Francesco Mambrini
Marco Passarotti
Caroline Sporleder

Proposed BibTeX entries:

```
@Proceedings{crh-2,
  title  = {Proceedings of the
      Second Workshop on        Corpus-
Based Research            in the
Humanities {CRH-2}},
  year   = {2018},
  editor = {Andrew U. Frank and
      Christine Ivanovic and
      Francesco Mambrini and
      Marco Passarotti and
      Caroline Sporleder},
  volume = {1},
  series = {Gerastree proceedings},
  isbn = {978-3-901716-43-0},
}

@InProceedings{crh2intro2018,
  author    = {Francesco Mambrini and Marco Passarotti
             and Caroline Sporleder},
  title     = {Preface},
  booktitle = {Proceedings of the Second Workshop on
                    Corpus-Based Research in the
                    Humanities {CRH-2}},
  year      = {2018},
  editor    = {Andrew U. Frank and Christine Ivanovic
             and Francesco Mambrini and Marco
             Passarotti and Caroline Sporleder},
  volume    = {1},
  series    = {Gerastree proceedings},
  pages     = {I-IV},
  isbn      = {978-3-901716-43-0},
}
```

ISBN 978-3-901716-43-0
Published by Gerastree Proceedings, GTP 1.
Dept. of Geoinformation, TU Wien, Austria.

Cover:
Les Fourches, seen from the Bretagne coast near Plouarzel (France). Photo by Andrew U. Frank.

Preface

The second edition of the international workshop on "Corpus-based Research in the Humanities" (CRH) is held in Vienna, hosted by Academy Corpora of the Austrian Academy of Science (https://www.oeaw.ac.at/ac/). It follows, on a biannual basis, the edition held in Warsaw in December 2015. But the origins of the workshop go back even further, CRH being the direct descendant of the former workshop on "Annotation of Corpora for Research in the Humanities" (ACRH), which was held three times: in Heidelberg (January 2012), Lisbon (November 2012), and Sofia (December 2013).

All the previous editions of ACRH/CRH were co-located with the international workshop on "Treebanks and Linguistic Theories" (TLT). This year, for the first time, CRH is an event on its own and it spans over two days. However, both the organizers of TLT and CRH worked to keep the connection between the two workshops as tight as possible and the two events as close as possible, both in time and place. As the sixteenth edition of TLT takes place in Prague in the two days before CRH, we hope that many scholars will be able to attend both workshops in a row. We want to thank very much Jan Hajič, the co-chairs of TLT-16 and the colleagues at the Institute of Formal and Applied Linguistics in Prague for doing their best to organize TLT in those days. And we thank Erhard Hinrichs, who took care of TLT since its first edition, for supporting ACRH/CRH as the co-located event of TLT for all these years.

Although for the 2015 edition of CRH we had received a rather limited number of submissions (17), we had the feeling that the topic of CRH was not only well motivated but also promising.

During the days in Warsaw, we met Andrew Frank and Christine Ivanovic from Vienna, who gave a joint talk there and were very positively impressed by both the motivations and the results of the workshop. In light of the ever growing Digital Humanities in Vienna, they offered to organize an edition of CRH there. We thought that this was a wonderful opportunity for CRH to grow and finally become independent. We accepted the invitation gladly and we are now happy to welcome Andrew Frank and Christine Ivanovic in the team of CRH's organizers as co-editors of these proceedings.

At Andrew Frank's suggestion, we selected "Time and Space Annotation" as the special topic for the Vienna edition of CRH. We believe the theme is aptly chosen, given the interest in the topic in the Viennese institutions and the international reputation of Andrew in the field. But in particular, we believe that the topic suits the aim of our workshop perfectly, especially in a period when geodata are becoming easier to access and increasingly dominant in many disciplines, while many fields experience what it is sometimes referred to as a "geographic turn". Spatial information is often used as linking point between different data sources, and gazetteers are adopted in the context of Linked Open Data. While time gazetteer are arguably still less mature, the interest in solutions that could use a grid of time-space coordinates for comparable Linked Open Data approaches is constantly raising. Potentially, chronological and spatial information in large corpora is a subject where research in archaeology, history, computational linguistics as well as ontologies and the semantic web can fruitfully converge.

Our hopes were fulfilled by a number of submissions, much higher than we expected. This year, we received 54 long abstracts (up to 6 pages) from scholars of 20 different countries all over the world: Austria, Brazil, China, Czech Republic, France, Germany, Greece, Hungary, India, Italy, Lithuania, Moldova, Norway, Portugal, Romania, Russia, Spain, Taiwan, UK and USA. We accepted 25 proposals, which corresponds to an acceptance rate of 46.3. The authors of the accepted abstracts were invited to submit full papers (up to 10 pages), which are collected in these proceedings. In the program, 18 proposals were presented as talks in oral sessions, while 7 made the poster session of the workshop.

Each submitted abstract was reviewed in double-blind fashion by three members of a program committee consisting of 36 scholars from 11 countries.

The program was completed by two invited talks (whose abstracts are published here), which tackle the special topic of this edition of CRH from different perspectives. The contribution by James Pustejovsky (Brandeis University, USA) focuses on semantic data modeling for temporal and spatial information from multimodal corpora, while Tara Andrews (University of Vienna) discusses a number of challenges opened by time and place annotation of historical data.

Looking at the table of contents of these proceedings, it becomes obvious that "Time and Space" was a felicitous choice as a special topic of this edition of CRH. 10 papers out of 25 deal with issues related with time and/or spacial information in corpora. 4 out of them focus on annotation questions, namely those by Ainara Estarrona and Izaskun Aldezabal (*Towards a Spatial Annotation Scheme for Basque based on ISO-Space*), by Katharina Korecky-Kröll and Lisa Buchegger (*Tagging spatial and temporal PPs with two-way prepositions in adult-child and adult-adult conversation in German in Austria*), by Dmitri Sitchinava and Boris Orekhov (*The Poetic Corpus of Russian: Where the Poems are Written*) and by Matthias Lindemann and Thierry Declerck (*Annotation and Classification of Locations in Folktales*). As it can be seen, 2 of them deal with the topic in literary or narrative texts.

Annotation is strictly linked both with its exploitation and with tools for automatic processing of data.

As for the former, the paper by Marie Mikulová, Eduard Bejček and Jarmila Panevová (*What Can We Find Out about Time and Space in ForFun Database?*) makes use of time and space annotation for investigating some linguistic phenomena of Czech, while those by Maria Moritz (*Time Proximity as a Means to Align Spelling Variants in old English Bibles: A Case Study*) and by Jean-Baptiste Camps (*Manuscripts in Time and Space: Experiments in Scriptometrics on an Old French Corpus*) are good examples of the exploitation of such annotation in the philological area. The paper by Venumadhav Kattagoni and Navjyoti Singh (*Towards an unsupervised learning method to generate international political event data using spatio-temporal annotations*) makes of use of metadata from time and space annotation for generating political events.

As for the latter, the papers by Adrien Barbaresi (*Towards a toolkit for toponym analysis in historical texts*) and by Delphine Bernhard, Pierre Magistry, Anne-Laure Ligozat and Sophie Rosset (*Resources and Methods for the Automatic Recognition of Place Names in Alsatian*) present present practical applications of tools and methods for automatic place annotation.

After time and place annotation, the main topic of this edition of CRH (like for the previous ones) is linguistic annotation of (mostly historical) corpora, covering quite diverse issues, ranging from annotation of poetry or legal texts to questions of normalization and dialect.

Given the wide variety of approaches and perspectives represented in the proceedings, we think that the area dealing with the use of empirical evidence provided by corpora for research in the Humanities is lively and diverse. Such diversity can be at the same time a pro and a con.

On one side, it allows to join different competences and research objectives on common issues. On the other hand, it runs the risk of missing a distinctive identity, which is essential to move from being just an empirical methodology to becoming a clear-cut research field. In this respect, the core issue is understanding what we mean with "the Humanities", at least in the CRH context. A tentative answer can come from looking at the papers published in these proceedings, which mostly deal with peculiar kinds of textual data stored in corpora, deviating from "regular" collections of linguistic empirical evidence, which want to include (supposedly) representative selections of modern languages. The corpora concerned in CRH papers feature texts in ancient/dead languages or diachronic varieties of modern ones, they feature either prose or poetry literary texts, and they open different kinds of philological questions.

All this implies and involves a wide variety of new end-users both of the corpora themselves and of the results of the research work coming from their use. Users will no longer be only linguists and NLP professionals, but philologists, historical linguists, classicists and scholars in literature.

The dialogue between such actors is sometimes not straightforward. Still today, the Humanities suffer an unfortunate separation between the so-called "Traditional Humanities" and "Digital Humanities". Most likely, CRH would be considered a "Digital Humanities" event. But such separation is today simply meaningless. In a sense, all the Humanities are now (at least partly) digital and there is no research in the Humanities that is not (at least partly) traditional. These two sides of the same coin must collaborate, as the heritage of centuries of research in the Humanities can now be optimally exploited thanks to new technologies, methodologies and resources, among which are corpora. CRH wants to support and encourage such coming together of different research paradigms and, ultimately, render the distinction between "Digital" and "Traditional" Humanities superfluous and old-fashioned.

We hope you will enjoy the workshop and the proceedings. We wish to thank all authors who submitted papers, the members of the program committee, James Pustejovsky and Tara Andrews, the local organizers and in particular Hanno Biber, Andrew Frank and Christine Ivanovic, who made the Austrian edition of CRH possible.

The CRH Co-Chairs
Francesco Mambrini (Deutsches Archäologisches Institut, Berlin, Germany)
Marco Passarotti (Università Cattolica del Sacro Cuore, Milan, Italy)
Caroline Sporleder (University of Göttingen, Germany)

Program Committee

Chairs:
Francesco Mambrini (Deutsches Archäologisches Institut, Berlin, Germany)
Marco Passarotti (Università Cattolica del Sacro Cuore, Milan, Italy)
Caroline Sporleder (University of Göttingen, Germany)

Members:
John A. Bateman (Germany)
Gerhard Budin (Austria)
Giuseppe Celano (Germany)
Arianna Ciula (UK)
Giovanni Colavizza (Switzerland)
Marco Coniglio (Germany)
Maud Ehrmann (Switzerland)
Andrew U. Frank (Austria)
Emiliano Giovannetti (Italy)
Stefan Th. Gries (USA)
Dag Haug (Norway)
Leif Isaksen (UK)
Christine Ivanovic (Austria)
Mike Kestemont (Belgium)
Puneet Kishor (Germany)
Dimitrios Kokkinakis (Sweden)
Sandra Kübler (USA)
Werner Kuhn (USA)
Piroska Lendvai (Germany)
Eleonora Litta (Italy)
Yudong Liu (USA)
Melanie Malzahn (Austria)
Roland Meyer (Germany)
Willard McCarty (UK)
John Nerbonne (The Netherlands)
Julianne Nyhan (UK)
Michael Piotrowski (Switzerland)
Geoffrey Rockwell (Canada)
Matteo Romanello (Germany)
Rainer Simon (Austria)
Neel Smith (USA)
Uwe Springmann (Germany)
Martin Thiering (Germany)
Sara Tonelli (Italy)
Martin Wynne (UK)
Amir Zeldes (USA)

Organising Committee

Chairs:
Francesco Mambrini (Deutsches Archäologisches Institut, Berlin, Germany)
Marco Passarotti (Università Cattolica del Sacro Cuore, Milan, Italy)
Caroline Sporleder (University of Göttingen, Germany)

Local Committee:
Hanno Biber
Andreas Dittrich
Andrew U. Frank
Katharina Godler
Christine Ivanovic

Contents

Incorporating Hittite into PROIEL: a pilot project

Guglielmo Inglese[1], Maria Molina[2] and Hanne Eckhoff[3]

[1]University of Pavia/University of Bergamo
[2]Institute of Linguistics, Russian Academy of Science
[3]Dep. of Modern Languages, University of Oxford
E-mail: guglielmo.inglese01@ateneopv.it; maria.lakhuti@gmail.com; hanne.eckhoff@mod-langs.ox.ac.uk

Abstract

In this paper, we report the results of a pilot project aimed at the inclusion of Hittite texts in the PROIEL family of treebanks. The first challenge is that the PROIEL annotation scheme has been designed for Indo-European languages mostly written in alphabetic scripts, so that a way to annotate the complex cuneiform script on which Hittite tablets are recorded must be worked out. Moreover, Hittite also provides some interesting morphosyntactic features that require the adaptation of annotation strategies already in use for other languages in PROIEL. Overall, our preliminary findings show that Hittite can be easily integrated into the PROIEL enterprise, but also that future work is required to effectively achieve this goal.

1 Introduction

The *Pragmatic Resources in Old Indo-European Languages* (PROIEL) project set out in 2008 with the aim of investigating information packaging and related phenomena, e.g. word order and discourse particles, in ancient Indo-European (IE) languages (Haug *et al.* [6]; Eckhoff *et al.* [2]). The core of the project consisted in the creation of annotated linguistic resources, i.e. treebanks, for the languages under analysis. In its earliest phase, the PROIEL corpus included the Greek text of the New Testament, along with its translations in Latin, Gothic, Old Church Slavonic (OCS), and Armenian (Haug *et al.* [5]). The texts were annotated in a layered scheme including lemmatization, morphological annotation, syntactic dependency annotation, and information structure.

Since its beginning, the PROIEL project has continuously grown and has nowadays become a standard for the annotation of ancient IE languages (Eckhoff *et al.* [2]). First, the treebanks of Greek, Latin, and Armenian texts have been expanded thanks to the addition of new textual material. Second, the PROIEL family of treebanks has been enriched with the addition of several newly created resources: the TOROT treebank, which includes Old Russian and OCS (Eckhoff & Berdicevskis [3]), the ISWOC treebank, featuring Old English, Old French, Old Spanish and Portuguese texts (Bech & Eide [1]), as well as new treebanks for ancient Germanic languages (see Eckhoff *et al.* [2] for details), namely Old Islandic (*Greinir skáldskapar*), Old Norwegian (*Menotec*), and Old Swedish (*MAÞIR*). Moreover, the treebanks featured in PROIEL have been recently converted to Universal Dependencies (UD).[1]

In spite of this positive trend of growth, there is still room for improvement, and the PROIEL project can be enhanced by the inclusion of additional IE languages. In this paper, we report on the results of a pilot project aimed at the integration of Hittite texts in PROIEL. Even though Hittite is the most anciently attested IE language and therefore of great interest for Indo-Europeanists, it remains a rather under-resourced language (Giusfredi [4]). First, reliable digital editions of Hittite texts are available only for a sub-set of the extant corpus (cf. *Hethitologie Portal Mainz*; https://www.hethport.uni-wuerzburg.de/). A linguistically annotated corpus is still a desideratum, even though this gap is progressively being filled: Inglese [10] laid out the basis for the annotation of Hittite texts according to the UD framework, with a focus on Old Hittite material, and a corpus of Middle and New Hittite material is currently being annotated with a constituency-based grammar at the project of the *Hittite Corpus* (HC; http://hittitecorpus.ru/; Molin & Molina [15]; Molina [16]). Therefore, adding Hittite texts to the PROIEL will not only improve the language coverage of the project, but will also considerably contribute to the creation of a much-needed digital resource in the field of Hittitology.

The paper is organized as follows. In Section 2, we briefly sketch the outline of the project and the material employed. We also discuss important issues connected with the preparation of the texts for the annotation and the philological issues that one needs to be aware of before digitizing Hittite texts. Section 3 contains an overview of the main problems encountered in the linguistic annotation of Hittite texts following the PROIEL's guidelines on different levels: tokenization (3.1), lemmatization and morphology (3.2), and syntax (3.3). Also, we briefly touch upon the crucial issue of fragmentary texts (3.4). We summarize our conclusions in section 4.

2 The Hittite pilot project

The pilot project was carried out in July and August 2016 and focused on the adaptation of the existing PROIEL annotation scheme to the necessities of Hittite. As a pilot study, we worked on the annotation of three Hittite texts: two New Hittite letters (KUB 19.5 + KBo 19.79, KUB 14.3, ed. by Hoffner [8]) and one Old Hittite instruction text (KBo 22.1, ed. by Miller [14]), for a total of 108 sentences. The annotation was manually performed by two

[1] See the project website for details (http://universaldependencies.org/). One of the anonymous reviewers asked why we opted for the PROIEL annotation scheme rather than annotating our data directly in the UD format. The reason is two-fold. On the one hand, PROIEL provides a more detailed scheme, in which we can include more structural information. Also, it allows the annotation of the semantic and the pragmatic layer, which is currently unavailable in UD. Another advantage is that the PROIEL scheme is stable, while UD is to some extent a moving target, as the scheme is still under considerable restructuring.

independent annotators (Maria Molina and Guglielmo Inglese) by means of the PROIEL Annotator web interface (Eckhoff *et al.* [2]) and the results and the issues that emerged during the annotation process were subject to extensive group discussions.

2.1 Material employed

As remarked in section 1, there is still a substantial lack of comprehensive digital editions of Hittite texts, and philologically reliable editions are mostly scattered across different sources. Clearly, this seriously hampers the possibility to carry out in-depth corpus analyses of the language and constitutes a further stimulus for the creation of a well-structured treebank of Hittite. However, unlike languages currently featured in the PROIEL, for which "the availability of electronic editions […] is relatively good" (Haug *et al.* [5]: 58), in the long run the inclusion of Hittite in PROIEL will require a good deal of manual digitalization of Hittite texts.

Texts for our pilot project have been kindly provided by Maria Molina from the HC, and are based on up-to-date philological editions. The texts were already split into sentences (see Molina [16] for the criteria behind sentence splitting; cf. Eckhoff *et al.* [2] on sentence splitting in PROIEL), and they were imported into the PROIEL annotation web interface by means of a script created by Hanne Eckhoff.

2.2 Text preparation: philological issues

The preparation of the texts for the annotation is not a trivial task, mostly owing to the philological complexity of the Hittite script. Unlike languages currently featured in PROIEL, such as Latin and Ancient Greek, which employ alphabetic scripts, Hittite is recorded in cuneiform script (see Hoffner & Melchert [7] for an overview), which poses several challenges for the digital annotation (Inglese [10]; Molina [16]).

The first issue is how to make the cuneiform script accessible to non-specialists of the language.[2] Two options are generally available: either texts are given in narrow transliteration, that is, each cuneiform sign is represented separately with hyphens as sign boundaries, as in *e-eš-zi* 'he is', or texts can be given in broad transcription, which is a rough phonological interpretation of the script, as in *ēšzi.* In our pilot, we have decided to give texts in broad transcription, which makes the corpus more readily available to users less acquainted with Hittite philology. However, it must be stressed that broad transcription requires a relatively high degree of normalization, so that most information about the cuneiform spelling is lost. Therefore, in the next steps of the project the narrow transliteration will be included in the corpus as well (see Inglese [10] for a possible solution), as it provides invaluable

[2] One of the anonymous reviewers asked why we have not decided to provide texts in cuneiform script with Unicode encoding. The reason is that texts in transcription are much more easily accessible to readers who have not been trained in Hittite philology. Moreover, Hittite cuneiform manuscripts are already digitized and freely available online at the *HPM*.

information on various linguistic facts, e.g. accent and vowel length, and spelling practices are worth investigating in their own right for various purposes (see e.g. Kloekhorst [13]).

Another peculiar feature of Hittite texts is that beside ‘syllabic’ signs, which stand for syllables in words written in Hittite and are commonly transliterated in lowercase italics, one also finds ‘logographic’ signs, i.e. sings which are read as Akkadian or Sumerian words, and are used as shortcuts for underlying Hittite words. As we discuss below, the annotation of Akkadograms and Sumerograms constitutes a remarkably tricky task. In addition, some Sumerograms, which are labelled ‘determinatives’, were graphically preposed to nouns to indicate the semantic class that a given noun belongs to.

To give an example of the complexity of the Hittite script, consider the passage in example (1), given in narrow transliteration. In this sentence, only the finite verb *ḫ*[*e*]*kta* ‘he bows’ is written in Hittite syllabic signs. As for the rest, one finds e.g. the Sumerian logograms LÚ standing for the Hittite nominative form *pešnaš* ‘man’, and the combination of the Akkadian preposition *ANA* ‘to’ with the Sumerogram LUGAL, which together stand for the Hittite dative form *ḫassui* ‘to the king’. Moreover, the determinative sign ^{d} preposed to the Sumerogram IM ‘Storm God’ indicates that the name refers to a deity.

(1) LÚ ^{d}IM *A-NA* LUGAL *ḫ*[*é-e*]*k-ta*
man storm.god to king bow.PRS.3SG.MID
“The man of the Storm God bows in the presence of the king.” (KBo 20.10 + KBo 25.59 i 5)

3 Linguistic annotation

In this section, we illustrate the main problems that we encountered in the linguistic annotation of Hittite texts following the PROIEL scheme. We discuss each layer of annotation separately, and highlight the most problematic issues. Notably, in the pilot the pragmatic level was left out.

3.1 Tokenization

Hittite scribes separated words through blank spaces, so that tokenization is a relatively trivial task. Still, some minor issues emerged in the course of the project. The first issue is how to tokenize and represent clitic chains in Wackernagel’s position (P2), which constitute a rarity among IE languages, but are systematic in Hittite. For instance, the graphic word *nu-wa-aš-ša-an* should be split up as *nu*=*wa*=*šan*, i.e. the sequence of the sentence initial connective *nu* plus the quotative particle =*wa* and the local particle =*šan*. For now, we have treated each item in the clitic chain as an independent token, and merely added the = sign to visually indicate token boundaries (cf. Eckhoff *et al.* [2] for the treatment of clitics in Old Portuguese in PROIEL). This is however a provisional solution, as one ideally needs a way to automatically retrieve whether a given token is a clitic or not. This might be achieved by inserting a dedicated tag at the morphological level.

The tokenization of determinative signs constitutes a further issue. In principle, determinatives can be tokenized either as distinct tokens or as a word-feature (for the discussion of pros and cons of both approaches see Inglese [10]). In our pilot, we have consistently adopted the former option, and treated determinatives in the same way as articles in Ancient Greek (see the treatment of LÚ.MEŠ in Fig. 2, sec. 3.3). It is unclear whether this strategy will be effective in the long run, and the annotation of more material is needed to gain a full appreciation of the issue.

Finally, another problem that was encountered is the treatment of Sumerian and Akkadian multi-word expressions, such as LÚ $^{\text{GIŠ}}$BANŠUR 'table attendant' and $^{\text{d}}$UTU=*ŠI* 'his majesty', which stand for single Hittite lexemes but are formally made up of multiple tokens in the languages they are written in. For the time being, we have resorted to annotating each token individually and indicating on the syntactic level that the two words belong to a single multi-word expression.

3.2 Lemmatization and morphology

Based on our pilot experience, the lemmatization and the annotation of morphological features are the layers of annotation that require the least adaptation of the existing PROIEL scheme.

Concerning lemmatization, for the sake of uniformity we have decided to give lemmas according to Tischler's glossary [12]. For most words, the stem form is used as the lemma, whereas for suppletive forms and *-r/n-*alternating stems the nominative is used instead. As common practice in PROIEL (cf. Eckhoff *et al.* [2]), homophonous lemmas are distinguished by storing them with variant numbers, e.g. *iya-#1* 'make' vs. *iya-#2* 'march'.

As for the morphological annotation, the tagset of morphological features in use in the PROIEL scheme requires minor modifications only. On the one hand, some of the existing features are not needed and can be simply left out, Hittite being notoriously morphologically simpler than languages such as Ancient Greek and Latin. On the other hand, new features were required, e.g. the 'ergative' case for neuter nouns ending in *-anza* when they occur as the subject of a transitive verb.

Further consideration is required for the lemmatization and morphological analysis of logograms. In general, one should decide whether to annotate these tokens according to the features of their surface language or according to the hypothesized features of their Hittite underlying forms. For the lemmatization, this implies a choice between Akkadian/Sumerian and Hittite lemmas for logograms. As the PROIEL scheme allows for a single lemma for each token only, we provisionally employed Hittite lemmas whenever available, and Akkadian/Sumerian ones in the rest of the cases. In the long run, it is desirable to develop a system in which logograms can be assigned both their surface lemma and their underlying Hittite lemma.

The issue of the morphological annotation of logograms is more complex. Beside plural markers on nouns (e.g. MEŠ), Sumerograms tend not to show overt morphological features. Therefore, they can either be left untagged, or they can be annotated according to the morphological features of their putative underlying Hittite forms. The situation of Akkadian is more complex. Unlike Sumerian, Akkadian forms in Hittite texts display a wider range of inflectional features, and some of them do not match the Hittite underlying forms. A case in point is the gender of 3rd sg. possessive pronouns, as in Akkadian one finds a masculine/feminine gender distinction =*ŠU* 'his' and =*ŠA* 'her' that is unparalleled in Hittite.

In our pilot, we tried to annotate all logograms according to their underlying Hittite forms, but this proves an unsatisfactory solution, because it greatly limits the possibility to search for logograms and their features in the corpus. Further work is needed to develop a solution to this issue.

3.3 Syntax

In PROIEL, the syntactic annotation, which constitutes the core of the treebank, is based on a dependency-style grammar.[3] The scheme was developed for the annotation of ancient IE languages, and it is quite suitable to annotate the syntax of Hittite texts as well. Consider the annotation of the Hittite complex sentence in (2), exemplified in Figure 1.[4]

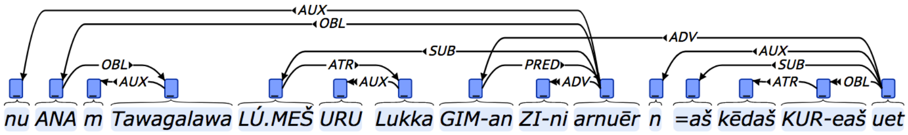

Figure 1: Annotation of a complex clause in Hittite

(2) [*nu*] *ANA* ^{m}*Tawagalawa* LÚ$^{\text{MEŠ}}$ URU *Lukka* G[IM]-*an*
CONN to *T.* man(PL) city *L.* when
ZI-*ni* [*a*]*rnuēr* *n*=*aš* *kēdaš*
soul.DAT bring.PST.3PL CONN=3SG.NOM DEM.DAT.PL
KUR-*eaš* *uet*
land.DAT.PL come.pst.1sg
"As the men of Lukka notified Tawagalawa, he came into these lands." (KUB XIV i 3-4)

Unsurprisingly, some minor modifications were required to allow for a more precise treatment of Hittite language-specific phenomena. In the first place, Hittite features various Wackernagel's (P2) clitic particles of partly unclear function, such as the so-called 'local particles', the connective

[3] See Eckhoff *et al.* [2] for a useful overview of the dependency grammar in use at PROIEL and the guidelines for details: <folk.uio.no/daghaug/syntactic_guidelines.pdf>.

[4] In this paper, Hittite dependency trees are visualized with *Arborator* (https://arborator.ilpga.fr/).

particles =*(m)a* and =*(y)a*, the quotative particle =*wa(r)*, and the particle =*za*. Following the PROIEL guidelines, since these items fail to show a syntactic function with respect to their head and loosely belong to the group of 'grammatical' words, we have consistently annotated them as AUX and assigned them a conventional head. However, we maintain that this annotation style is too opaque, as it does not allow a sufficient differentiation between items bearing the AUX relation. A more fine-grained tagset should be worked out. Similarly, we also annotate preverbs, which are never univerbated with the verbal stem they modify, as AUX, as in the case of *anda* 'in' in Fig. 2 below.

Another construction which deserves more attention is the relative clause. So far, relative clauses in PROIEL have been treated as embedded predications depending on a noun, and correlative relative clauses, which marginally occur in e.g. Latin, do not receive a dedicated annotation.

However, there is evidence that correlative relative clauses are not syntactically part of the main clause, as they do not modify an external head noun, nor can they fill in the valency frame of a predicate (cf. Inglese [11] with further references). As correlative clauses constitute the default relativization strategy in Hittite, we have devised a new annotation style to capture the linguistic reality of this phenomenon. In our scheme, the verb of the correlative clause depends on the verb of the main clause via the newly created *rel* tag. As an example, consider the annotation of the sentence in (3) given in Figure 2.

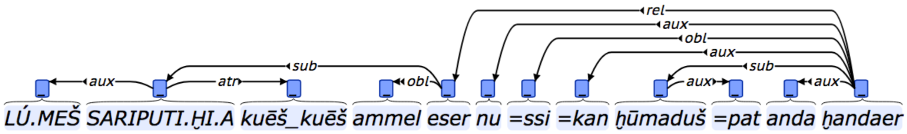

Figure 2: Annotation of correlative clauses

(3) $^{\text{[LÚ].MEŠ}}$*SARIPUTI*$^{\text{ḪI.A}}$ *kuēš* *kuēš* *ammel*
purple-dyer(PL) REL.NOM.PL REL.NOM.PL 1SG.GEN
eser [*nu=ssi=kan* *ḫ*]*ūmaduš=pat* *anda*
be.PST.3PL CONN=3SG.DAT=PTC all.NOM.PL=FOC in
ḫandaer
align.PST.3PL
"All the purple-dyers who were mine, they all joined him." (KUB 19.5 + 10)

Finally, Hittite features different periphrastic constructions, or compound verb forms. In these cases, we follow PROIEL's approach and treat constructions with *ḫark-* 'have' and *eš-* 'be' plus participle as grammaticalized monoclausal constructions when they show a perfect or a passive reading (for discussion see Hoffner & Melchert [7]; Inglese & Luraghi [9]). The annotation of a perfect with *ḫark-*, quoted in (4), is exemplified in Figure 3. As the figure shows, the participle *ḫazzian* ‘pierced’ is treated as the head of the predication, whereas the finite verb *ḫarzi* ‘has’ is tagged as AUX.

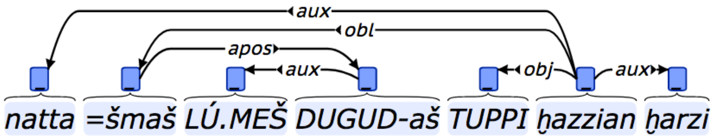

Figure 3: Annotation of the periphrastic perfect

(4) *natta=šmaš* $^{\text{LÚ.MEŠ}}$DUGUD-*aš* *TUPPI* *ḫazzian*
NEG=2PL.DAT dignitary.DAT.PL tablet pierce.PTCP.N/A.N
ḫarzi
have.PRS.3SG
“(As my father keeps writing to you), has he not written the tablet to you dignitaries?” (KBo 22.1 i 23)

Conversely, in the case of the ‘stative’ *ḫark-* and *eš-* plus participle and the ‘ingressive’ *dai-/tiya-* ‘put’ plus supine constructions, we take the finite verb as the head of the predication, and tag the accompanying verb as XOBJ. As an example of the treatment of the stative construction, consider the annotation of example (5) in Figure 4. The finite verb *ḫarzi* is the root of the tree, and the participle *tamaššan* ‘oppressed’ depends on it as XOBJ.

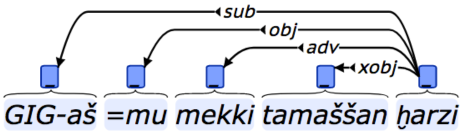

Figure 4: Annotation of the ‘stative’ periphrastic construction

(5) GIG-*aš=mu* [*mekki*] *tamaššan* *ḫarzi*
illness.NOM=1SG.ACC much oppress.PTCP.N/A.N have.PRS.3SG
“Illness keeps me severely prostrated.” (KUB 19.5 + i 5-6)

Notably, the ‘serial’ constructions with *pai-* ‘go’ and *uwa-* ‘come’ do not easily fit in either scheme: since we did not encounter them in out pilot, we leave the design of an appropriate annotation style for the future

3.4 Fragmentary texts

Another crucial issue concerns the annotation of fragmentary sentences, i.e. sentences which are only partly readable because of the poor conservation status of the manuscript. As discussed at length by Molin & Molina [15] and Inglese [12], various options are available for the annotation of fragmentary sentences. Given the complexity of the topic, in the pilot we have avoided the annotation of such sentences. In principle, we aim at an annotation halfway between what suggested by Molin & Molina [15] and

Inglese [12]: sentences should receive a sentence tag according to their 'brokenness' level, together with a more fine-grained sentence internal annotation of philological gaps as tokens. In the future, this will require a special adaptation of the PROIEL scheme, which so far does not allow the tagging of features at the sentence level.[5]

4 Conclusions and future work

In this paper, we have reported on the preliminary results of a pilot project aimed at the inclusion of Hittite into the PROIEL enterprise. This is a much needed and welcome expansion of the resource. On the one hand, it will enrich the current language coverage of the PROIEL project, while at the same time ensuring the creation of the first dependency-based treebank for Hittite. We have shown that the PROIEL guidelines by and large easily lend themselves to the annotation of Hittite. However, Hittite texts presents several philological difficulties which requires further consideration, in order to provide a reliable and user-friendly digital resource. Finally, we have also discussed how the guidelines should be partly tailored to annotate a number of language-specific constructions of Hittite. Overall, our findings provide the necessary starting point for the creation of a Hittite treebank within the PROIEL framework.

References

[1] Bech, K. & Eide, K. 2014. *The ISWOC corpus*. Department of Literature, Area Studies and European Languages, Oslo.

[2] Eckhoff, H. M., Bech, K., Bouma, G., Eide, K., Haug, D. T. T., Haugen, O. E., Jøndal, M. 2017. The PROIEL treebank family: a standard for early attestations of Indo-European languages. *Language Resources and Evaluation*, https://doi.org/10.1007/s10579-017-9388-5.

[3] Eckhoff, H. M. & Berdicevskis, A. 2015. Linguistics vs. digital editions: The Tromsø Old Russian and OCS Treebank. *Scripta & e-Scripta* 14–15: 9-25.

[4] Giusfredi, F. 2014. Web resources for Hittitology. *Bibliotheca Orientalis* 71: 358-362.

[5] Haug, D. T. T., Eckhoff, H. M., Majer, M., Welo, E. 2009. Breaking down and putting back together: analysis and synthesis of New Testament Greek. *Journal of Greek Linguistics* 9 (1): 56-92.

[5] As one of the anonymous reviewers suggests, fragmentary contexts might be handled in a similar way to disfluency phenomena in treebanks of spoken data (see e.g. the guidelines for the annotation of disfluency in UD). This similarity is based on the insight that the two phenomena involve the unexpected disruption of the 'natural' syntax of a sentence. However, we suspect that fragmentary contexts are more varied and complex to annotate than disfluencies, so that further work is needed to develop an *ad hoc* solution.

[6] Haug, D. T. T., Jøhndal, M., Eckhoff, H. M., Welo, E., Hertzenberg, M. J. B., & Müth, A. 2009. Computational and Linguistic Issues in Designing a Syntactically Annotated Parallel Corpus of Indo-European Languages. *Traitement Automatique des Langues* 50 (2): 17-45.

[7] Hoffner, H. A. & Melchert, C. H. 2008. *A Grammar of the Hittite Language. Part I: reference grammar.* Winona Lake (Indiana): Eisenbrauns.

[8] Hoffner, H. A. 2009. *Letters from the Hittite Kingdom.* Atlanta: Society of Biblical Literature.

[9] Inglese, G. & Luraghi, S. Forthcoming. The Hittite Periphrastic Perfect. To appear in *Perfects in Indo-European languages*, vol. I, R. Crellin & T. Jügel (eds.). Amsterdam/Philadelphia: John Benjamins.

[10] Inglese, G. 2015. Towards a Hittite Treebank. Basic Challenges and Methodological Remarks. In *Proceedings of the Workshop on Corpus-Based Research in the Humanities*, Passarotti, M., Mambrini, F., & Sporleder, C. (eds.), 59-68.

[11] Inglese, G. 2016. La classificazione delle frasi relative in ittita arcaico: una prospettiva tipologica. *Studi e Saggi Linguistici* 54: 9-44.

[12] Inglese, G. Forthcoming. Annotating the syntax of fragmentary sentences: the case of Hittite. To appear in *Formal Representation and Digital Humanities*, P. Cotticelli & F. Giusfredi (eds.). Cambridge: Cambridge Scholars Publishing.

[13] Kloekhorst, A. 2014. *Accent in Hittite: A Study in Plene Spelling, Consonant Gradation, Clitics, and Metrics*. Wiesbaden: Harrassowitz.

[14] Miller, J .2013. *Royal Hittite Instructions and Related Administrative texts.* Atlanta: Society of Biblical Literature.

[15] Molina, M. & Molin, A. 2016. In a Lacuna: building a syntactically annotated corpus for a dead cuneiform language (on the basis of Hittite). In *Proceedings of the International Conference "Dialogue 2016".*

[16] Molina, M. 2016. Syntactic Annotation for a Hittite Corpus: Problems and Principles. In *Proceedings of the Workshop on Computational Linguistics and Language Science*.

[17] Tischler, J. 2001. *Hethitisches Handwörterbuch*. Innsbruck: Institut für Sprachwissenschaft.