Phrase Structure and Ancient Anatolian languages
Methodology and challenges for a Luwian syntactic annotation

Federico Giusfredi
Dipartimento di Filologia, Letteratura e Linguistica
University of Verona
federico.giusfredi@gmail.com

Abstract

English For the Marie Skłodowska Curie (MSCA) funded project "SLUW – A computer aided study of the (morpho)-syntax of Luwian" a collection of phrase structure trees from the Luwian corpus is currently being prepared. Luwian is a language belonging to the Anatolian branch of Indo-European; its structures are different from those of English and the language itself is partly obscure. The present paper will describe some special needs, open challenges and methodologies relevant for the annotation of phrase-structure of Luwian.

Italiano *Per il progetto Marie Skłodowska Curie "SLUW – A computer aided study of the (morpho)-syntax of Luwian", è in preparazione un'ampia collezione di alberi sintattici a costituenti per il corpus luvio. Il luvio era una lingua del ceppo anatolico dell'indoeuropeo; la sua struttura è diversa da quella dell'inglese, e la sua decifrazione è in parte incompleta. In questo articolo, saranno discusse alcune necessità, problemi e metodi rilevanti per l'annotazione della sintassi dei costituenti del luvio.*

1 Introduction

Annotating a dead language, especially if lacunae and obscure sequences occur frequently in the corpus, is a challenging task. In the case of phrase-structure trees, those challenges complicate the usual issues represented by "trapping" (an element nested within the boundaries of a phrase it does not belong to) and standard discontinuous phrases.

The language under investigation is Luwian, an ancient member of the Anatolian branch of Indo-European, the second largest one after Hittite by number of documents. It was written using two different writing systems (the cuneiform script and the Anatolian hieroglyphs). The attestations cover a time span of almost one millennium, between the 16^{th} and the 8^{th} centuries BCE (cf. Melchert, 2003).

Syntactically speaking, it features a rather strict SOV word-order as far as some classes of constituents are concerned (Wackernagel particles, inflected verb at the end, left-branching of genitives and attributes); while a few elements can move with relative freedom (for instance adverbs, indirect case NPs and PPs with respect to the position of a direct object).

The final goal of the SLUW project, a Horizon2020 MSCA funded two-year research plan hosted by the University of Verona (2015-2017) is to produce a general study of the syntax (and morpho-syntax) of the language; in order to do so, a significant selection of sentences (about 30% to 50% of the corpus) will be collected and annotated in order to produce phrase-structure trees that will help highlight syntactic patterns. Theory-free phrase structure annotation is more suitable than Universal Dependencies for this kind of approach, as the boundaries of linear and non-linear phrases as well as their canonical or non-canonical position within the sentence are more easily identified.

Since the structure of Luwian is very different from the one of English – Anatolian languages had peculiar features that must be accounted for – the starting point for the development of a POS tagset, the "label-tag" context-sensitive system of the Penn Treebank II, requires to be modified in order to better match the object of study.

2 Expanding the tagset

Different languages have different features, and some of them may be especially relevant for the understanding of the syntax (or of any other aspects of its nature that may be of interest). In the case of Luwian, the Penn POS system (Taylor, Marcus and Santorini, 2003) needs to be expanded on both the phrase and the word level. The following addenda represent the state of the Luwian tagset as of September 2015; other modifications will certainly occur during the future analysis of the corpus.

On the phrase level, the preliminary analysis indicated that the following elements need to be added to the POS labels:

CLP	Clitic "Phrase"
INTR	Introductory particle
QUOT	Direct speech marker

CLP is a pseudo-node (it does not represent a real constituent). In Luwian, a large set of particles with different functions is bound to P2 (2nd word position) – some belonging to the VP, some working on the sentence or inter-phrasal level. While "movement" may be assumed for argumental elements, a proper analysis of some of these clitics has not yet been attempted. They will therefore be analyzed in the position that they actually occupy in the phrase structure, at least during the theory-free phase of annotation.

INTR is a typical element of the Anatolian syntax: an accented particle that works as a coordinating conjunction, but may also open any sentence in which no other accented elements occur before the Wackernagel particles.

Finally, QUOT is a direct speech marker that quite frequently occurs in Wackernagel position.

On the word level, most of the special features of the Anatolian languages can be dealt with by wisely using a functional architecture (matching case endings, verbal inflection; cfr. Taylor, Marcus and Santorini, 2003; also Marcus et al., 1994). Formal markers for nominal elements will include case(-like) specifiers, such as:

-NOM	Nominative
-ACC	Accusative
-GEN	Genitive
-DAT	Dative
-ABL	Ablative
-VOC	Vocative
-NAN	Nom./Acc. (neutra)
-ANT	*-ant-* form (ergative-like)

For verbs, marking endings, time, mood, and voice is also of the utmost importance:

-#S/P	#th person singular/plural
(-)T	Past tense
(-)I	Imperative
(-)MP	Medio-Passive

The case-attributes are important because simply co-indexing elements belonging to the same phrase would make it difficult to assess the cases in which the agreement between two or more elements is not perfect.

This happens in some cases with certain Anatolian modifiers (numerals and nouns do not always agree in number) and with some types of syntactic alignment ("ergative"-like *ant*-forms are modified by attributes in common-gender nominative, and can be anaphorically recalled by neutral pronouns).

Apart from these functional tags, on the word-level specific POS tags also need to be added. For instance, as far as adjectives are concerned:

GJJ	Genitival adjective
PJJ	Possessive adjective
REL	Relative "pronoun"

GJJ represents a peculiar type of synchronically productive adjective that was used to replace the genitive case (cf. Bauer, 2014, 147ff.), an example being *mayas(s)a/i-* "of the adult(s)". It implied a genitival relationship to *maya-* "adult"; it was inflected and agreed with the regens, thus we may have ablative (instrumental):

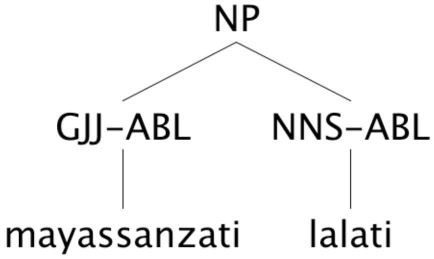
[1] *mayassanzati lalati*

adult=gen.adj.=pl.=abl. tongue=abl.

"The tongues of the adults"

(text KUB 35.24 i 4)

which results in the constituent-structure represented in the following tree.



In case of more complex genitival chains, the nesting of the constituents disambiguates different levels of possession, for instance:

[2] *sasaliya Maritis Zwarimis* FILIUS-*muwiyaya*

sasali=n/a=pl. PN$_1$=gen. PN$_2$=gen. son=gen.adj.

sasali's of Maritis, son of Zwarimis

(text Malatya 3, §1)

Tags must therefore be available in order to mark the structure of the phrases and disambiguate from other genitival strategies. PJJ are possessive adjectives similar to English *my*, but they also require inflection and agreement, as in the case of GJJ.

2.1 Subordination and relative clauses

A preliminary analysis has shown that, in some cases, Anatolian subordinate clauses contain a complex set of candidate "nodes" on the level of the SBAR element of the POS tagset, that would roughly correspond to the CP node of a transformational tree: the so-called Anatolian "connectives" (INTR) and subordinating conjunctions may co-occur, and this calls for caution as far as the syntactic representation is concerned.

Consider for instance the following example, in which the syntactic status of the first INTR-element *a* is problematic, because the "complementizer"-slot in the subordinate is already taken by the subordinating conjunction *kuman*, and the "complementizer"-slot of the main clause is occupied by another INTR-element, which makes the intepretation of the subordinate as embedded impossible (or at least very difficult).

[3] **[$_{\text{INTR}}$ a]** [$_{\text{S}}$ [$_{\text{SBAR}}$ **t-1** [$_{\text{QUOT}}$ wa] [$_{\text{VP}}$ [$_{\text{NP-OBJ}}$ kummaya DEUS.DOMUS-sa] **[$_{\text{IN-1}}$ kuman]** **[**$_{\text{V}}$ tamaha]]] **[$_{\text{INTR}}$ a** [$_{\text{QUOT}}$ wa] [$_{\text{NP}}$ mu] [$_{\text{PTCL}}$ tta] [$_{\text{VP}}$ [$_{\text{DP-SBJ}}$ zanzi kutassarinzi] [$_{\text{V}}$ appan awinta]]]]

"And, when I built the holy temples, these orthostats followed me."

(text Karkemish A11a §§14f.)

The identification of this problem (that also exists in Hittite) has important theoretical consequences regarding the inter-phrasal syntax of Anatolian: "connectives" like *a* were so far consistently presented as coordinating elements, but apparently this is not always the case (cf. Cotticelli-Kurras and Giusfredi 2015).

As for the REL label, the treatment of relative sentences in Anatolian is rather peculiar. The two clauses formally appear to be coordinated; the relative element in the relative clause is frequently referred to a nominal element (Hoffner and Melchert, 2008, 423-424). In such cases, it is inflected to agree with the noun, and is recalled by a pronoun in the main clause. A pseudo-English example can be the following:

[4] *___to what man___ _you spoke,_ ___that___ _is a liar_

what=dat. man=dat. you spoke, that=nom. is a liar

Therefore, the REL element needs to be assigned the range of attributes of an adjective.

3 Lacunae and *cruces*

Lacunae in a text preserved on a clay tablet – or on any other kind of perishable support – may interfere with the parsing of syntactic structures. So does the presence of segments or sequences of segments that have not been fully deciphered.

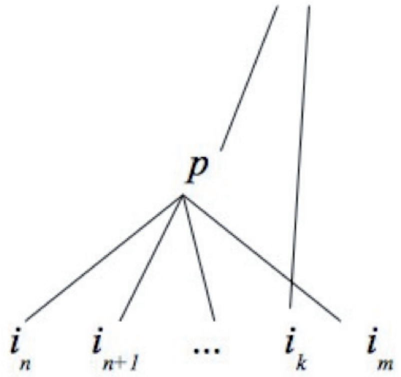
From the point of view of phrase-structure annotation, these two peculiarities of the corpora of ancient dead languages can occur in two different forms: either the unparsable element is an isolated node on the phrase level, or it belongs to a complex phrase, along with other elements that are analyzable.

In the first case, the unparsable element can simply be assigned a specific tag – in a way similar to the <damage> XML tag proposed by Korkiakangas and Lassila (2013). A similar problem has also been discussed by Zemánek (2007), in the framework of a treebank of the ancient Semitic Ugaritic language.

When, on the contrary, the unparsable element(s) interrupt(s) a phrase, the problem can be seen as a special case of phrase discontinuity (in other words, it is formally identical to the case in which a dislocation or movement produces discontinous phrases).

3.1 Discontinuous phrases

Discontinuous phrases, both the "sprachwirklich" ones and the ones produced by an unparsable element, can be formally defined as follow. Rephrasing the definition of yield Y of a node p given by Kallmeyer, Maier and Satta (2009; cf. Maier, 2011) as the set of all the indices $i \in \mathbb{N}$ such that p dominates the leaf labeled with the i^{th} terminal, one can generalize the definition of "discontinuous phrase" as follows. A phrase that is mapped at the node p with yield Y is a discontinuous phrase iff for $I_n \in Y \ \exists\, m > n$ such that $I_m \in Y$, $\exists\, k$ such that $n < k < m$ and $i_k \notin Y$.



Discontinuity can, in several cases, be solved employing iterations or recursive strategies; however, from the point of view of linguistic representation, this may, in given circumstances (such as trapping), interfere with the morphosyntactic notation (nesting NPs will not always solve the problem of a discontinuous NP containing an extraneous element such as a preverb).

In the cases where nesting is not a valid option, using attribute indexing and pointers (Taylor, Marcus and Santorini 2003) in order to co-index the components of a phrase (for a formal definition of component see Kallmeyer, Maier and Satta, 2009) appears to be the best strategy available.

4 Conclusion

The creation of phrase-structure trees for ancient languages with structural peculiarities that make them very different from modern ones may require specific modifications to the usual parsing tagsets. Such modifications may occur both on the phrase and on the word levels. In order to minimize the challenges and maximize flexibility, a context-sensitive syntax with both labels and functional tags is more suitable than a rigid one; for instance functional markers for case inflection may apply to several different categories of labels (all nouns, adjectives and pronouns).

As far as discontinuous phrases are concerned, in the analysis of dead languages they may be natural linguistic phenomena, but they may also be the result of either poor text preservation or limited understanding of given segments. In order to avoid inaccurate nesting, a system of co-indexing appears to be the most advisable solution to guarantee a good degree of accuracy in the linguistic representation and a regular treatment of the linearity issues.

References

Anna Bauer. 2014. *Morphosyntax of the Noun Phrase in Hieroglyphic Luwian*, Brill, Leiden.

Paola Cotticelli-Kurras and Federico Giusfredi. 2015. *On Luwian Syntax: presentation of the SLUW project*, paper presented at the Arbeitstagung der Indogermanischen Gesellschaft, Marburg, 21 September 2015.

J. David Hawkins. 2000. *Corpus of Hieroglyphic Luwian Inscriptions, Volume I, Inscriptions of the Iron Age*. De Gruyter, Berlin/New York.

Harry A. Hoffner and H. Craig Melchert. 2008. *A Grammar of the Hittite Language*. Brill, Leiden.

Laura Kallmeyer, Wolfgang Maier and Giorgio Satta. 2009. *Synchronous rewriting in treebanks*. Proceedings of the 11th International Conference on Parsing Technologies. Paris: 69-72.

Timo Korkiakangas and Matti Lassila. 2013. *Abbreviations, fragmentary words, formulaic language: treebanking mediaeval charter material*. Proceedings of The Third Workshop on Annotation of Corpora for Research in the Humanities. Sofia.

KUB 35 = *Keilschrifturkunden aus Boghazköi*, Band 35, 1993. Gebr. Mann, Berlin.

Wolfgang Maier. 2011. *Characterizing Discontinuity in Constituent Treebanks.* Formal Grammar Lecture Notes in Computer Science Volume 5591: pp 167-182

Mitchell Marcus, Grace Kim, Mary Ann Mrcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Jatz, Britta Schasberger. 1994. *The Penn Treebank: Annotating Predicate Argument Structure*. University of Pennsylvania, Philadelphia.

H. Craig Melchert. 2003. *The Luwians*. Brill, Leiden.

Ann Taylor, Mitchell Marcus and Beatrice Santorini. 2003. *The Penn Treebank: An Overview.* University of York. Heslington, York.

Petr Zemánek. 2007. *A Treebank of Ugaritic. Annotating Fragmentary Attested Languages*. Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories. Bergen.