Morpho-Syntactic Analyso-Parsing

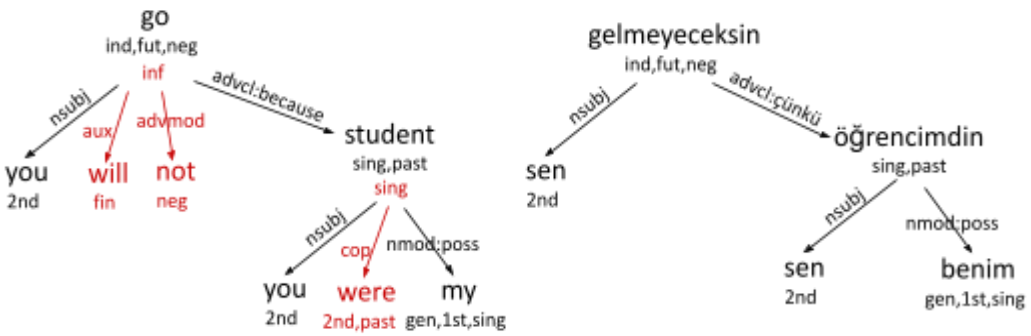
Omer Goldman, Leonie Weissweiler & Reut Tsarfaty

We suggest combining syntactic parsing with morphological analysis in a way that avoids (most) theoretical debates on word boundaries. Combined morpho-syntactic data will be more inclusive towards languages that are currently treated unnaturally – most prominently noun-incorporating languages. Combined morpho-syntactic models will be able to parse sentences in more languages and enable better cross-lingual studies.

The Data

The data will be based on existing treebanks and incorporate function words as phrase-level morphological features, leaving only content words as nodes of a dependency graph. The move from segmentation of words to a distinction between content and function could eliminate most issues regarding inconsistent segmentation, either across languages[1] or across treebanks of the same language.[2]

In isolating languages, the data will explicitly surface morpho-syntactic features that are expressed periphrastically. Below are trees of parallel sentences in English and Turkish as an example: *you will not go because you were my student* and *sen gelmeyeceksin çünkü sen benim öğrencimdin*. The English tree, with both its red and black nodes, is the current parse tree, while a word-free morpho-syntactic representation contains only the nodes in black. The nodes that correspond to function words are replaced with extra features on the parent node. Specifically, the auxiliary *will* and adverb *not* are not represented as independent nodes, but contribute to the features of *go*, and likewise for the copula *were* and its parent *student*. Phrase-level features avoid the need to decide what is a morpheme vs. an auxiliary.

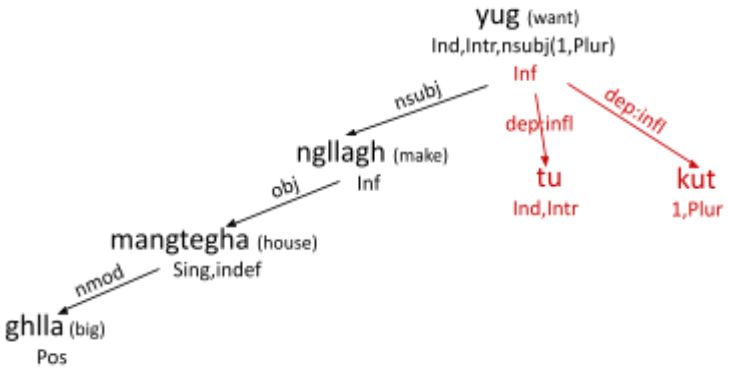


On the other hand, in polysynthetic languages, the addition of phrase level features to content words will expose the argument structure even if it is encapsulated in a single word. Below is the Yupik tree of the sentence *mangteghaghllangllaghyugtukut* (*we want to make a big house)*. Currently, noun-incorporating languages like Yupik undergo a full morpheme segmentation, resulting in agreement morphemes like *kut* and TAM markers like *tu* appearing as "words" in the parse tree. Assigning phrase-level features only to content lexemes will eliminate the function nodes (in red) and will make the Yupik morpho-syntactic parse tree similar to its English equivalent.

---

[1] E.g., Japanese is treated as isolating and Korean as agglutinative, although they are very similar typologically.

[2] E.g., the different treebanks for Hebrew segment and attribute different surface forms for clitics.



In practice, we suggest amending the CoNLL-U files to include phrase-level features for content words that will be annotated mostly automatically from the current word-level morphological features and the function words. The table for the English example sentence is given below. Including all nodes will result in a regular dependency graph, while ignoring all nodes without phrase-level features will result in a morpho-syntactic tree for this task. Our initial annotation effort in English and Hebrew found that most of the annotation could be done automatically using a grammar, although some manual decisions are to be taken.

ID	Form	Lemma	POS	FEATS	HEAD	DEPREL	P-FEATS
1	you	you	PRON	Nom,2,Sing	4	nsubj	Nom,2,Sing
2	will	will	AUX	Fin	4	aux	
3	not	not	PART	Neg	4	advmod	
4	go	go	VERB	Inf	0	root	Fin,Ind,Fut,Neg
5	because	because	SCONJ	-	9	mark	
6	you	you	PRON	Nom,2,Sing	9	nsubj	Nom,2,Sing
7	were	be	AUX	Fin,Ind,Past, 2,Sing	9	cop	
8	my	my	PRON	Gen,1,Sing	9	nmod:poss	Gen,1,Sing
9	student	student	NOUN	Sing	4	advcl:because	Sing,Ind,Past

The Task

Since function words do not appear as nodes in a morpho-syntactic tree, the phrase-level morphological features are essential in order to fully characterize a sentence. Thus, the task will require models to predict both the labeled dependency arcs and the morphological features. This would require a combination of standard parsing models that do not predict features, and morphological analysis models that only predict features.

The evaluation metrics are to be determined, but could be a combination of LAS for the arcs and accuracy for the features. The nature of the task will allow the inclusion of a diverse set of languages, both genealogical and typologically, more diverse than the usual selection in most parsing tasks.

Participants could model this task either as a combination of separate parsing and analysis models, or come up with novel models that solve the entire task in one stage. Successful models would be much more suitable for prediction of predicate-argument structure in polysynthetic languages, as well as for automatic or semi-automatic annotation of data in low-resourced languages. This will enable better cross-lingual studies, both due to the inclusion of more diverse languages and due to the results' independences from orthographic traditions.