Redefining Syntactic and Morphological Tasks for Typologically Diverse Languages

Omer Goldman, Leonie Weissweiler and Reut Tsarfaty

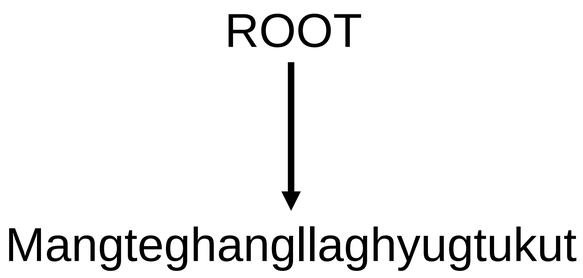
Motivation

NLP for polysynthetic languages! 🥳

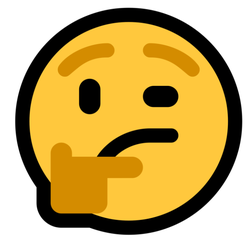
Let's try Yupik:
Mangteghangllaghyugtukut

English: We want to make a house

Dependency Parsing



Morphological Analysis



Word Segmentation

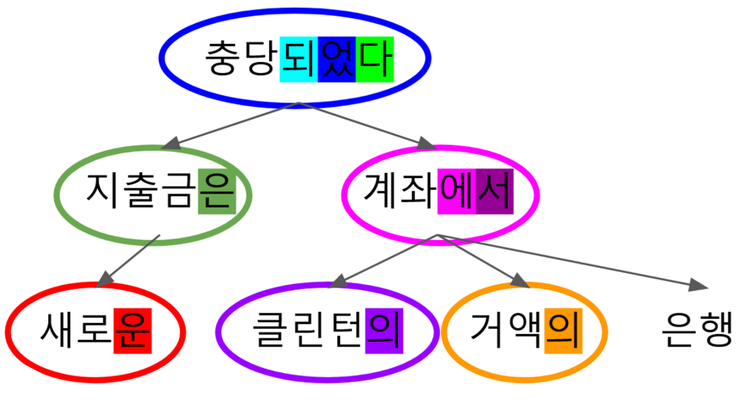
Inconsistent

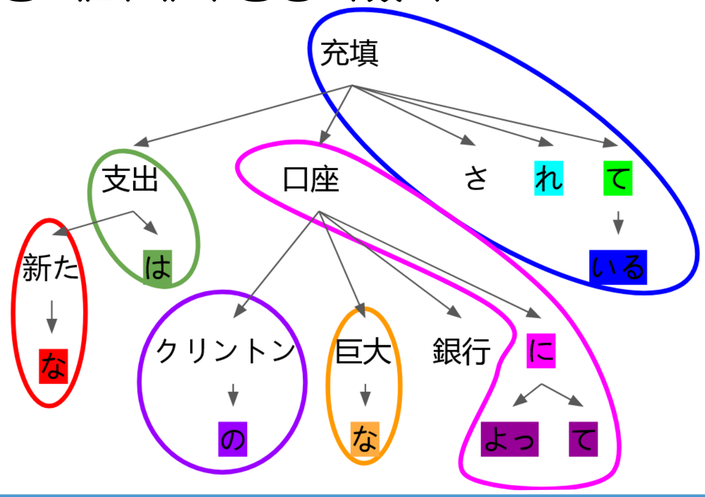
- across languages
- across treebanks

Hebrew: אהבתיה

- HTB: אהבתי -את -היא
- IAHLT: אהבתי -ה

English: The new spending is fueled by Clinton's large bank account.
Japanese: 新たな支出はクリントンの巨大な銀行口座によって充填されている。
Korean: 새로운 지출금은 클린턴의 거액의 은행 계좌에서 충당되었다.

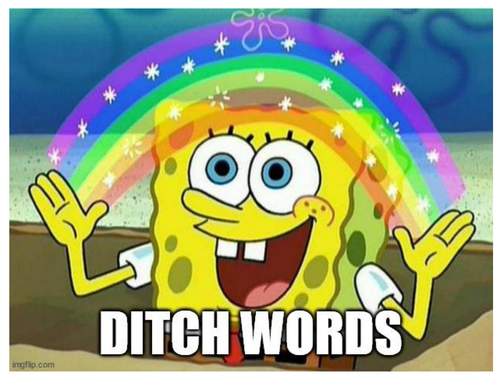




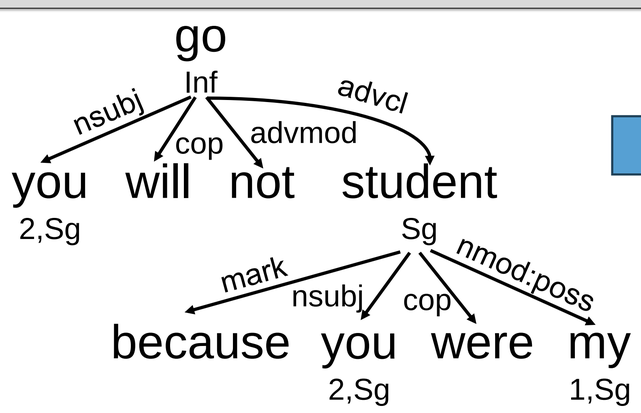
Our Solution

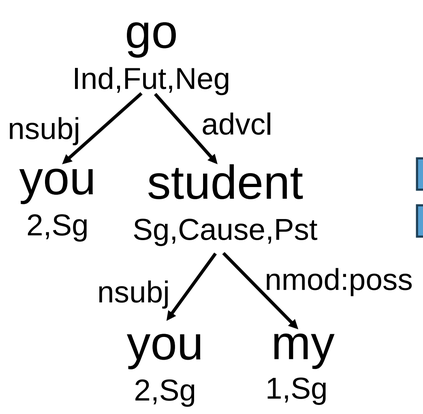
Content words/morphemes → "syntactic" nodes

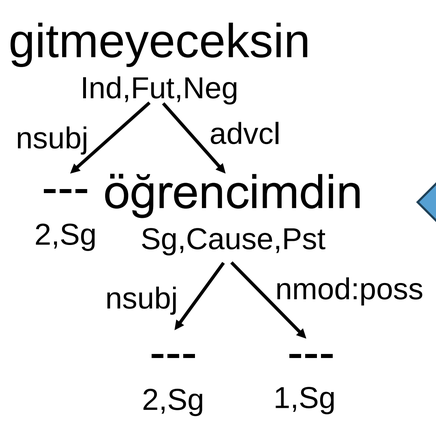
Function words/morphemes → "morphological" features

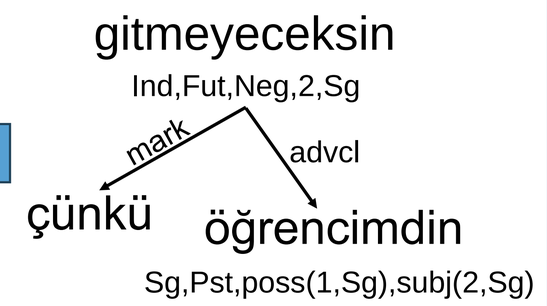


More parallel trees









UD-based Data

- Minimal differences from UD
- Nodes with MS-features → MS tree
- All nodes → dependency tree
  (When possible)

ID	Form	Lemma	POS	FEATS	HEAD	DEP	MS-FEATS
1	you	you	PRON	Nom;2;Sg	4	nsubj	Nom;2;Sg
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;Sg	9	nsubj	Nom;2;Sg
7	were	be	AUX	Fin;Ind;Past;2;Sg	9	cop	
8	my	my	PRON	Gen;1;Sg	9	nmod:poss	Gen;1;Sg
9	student	student	NOUN	Sg	4	advcl:because	Sg;Ind;Past