Universal dependencies for Uyghur

Mairehaba Aili
Xinjiang University,China
marhaba@xju.edu.cn

Weinila Mushajiang
Xinjiang University,China
winira@xju.edu.cn

Tuergen Yibulayin
Xinjiang University, China
turgun@xju.edu.cn

Kahaerjiang A.
Xinjiang University
kaharjan@xju.edu.cn

Yan Liu
Xinjiang University
liuyuxiu@xju.edu.cn

Abstract

The Universal Dependencies (UD) Project seeks to build a cross-lingual studies of treebanks, linguistic structures and parsing. Its goal is to create a set of multilingual harmonized treebanks that are designed according to a universal annotation scheme. In this paper, we report on the conversion of the Uyghur dependency treebank to a UD version of the treebank which we term the Uyghur Universal Dependency Treebank (UyDT). We present the mapping of the Uyghur dependency treebank's labelling scheme to the UD scheme, along with a clear description of the structural changes required in this conversion.

1 Introduction

Treebanks can be used for statistical learning as well as evaluation and are available for an increasing number of languages. For instances: Czech (Hajičová, 1998), Danish (Kromann, 2003), Turkish (Oflazer, 2003) Slovene (Džeroski et al., 2006), and Finnish (Haverinen et al., 2010). However, because of having been built with language-related specific schema, it leads to different treebanks with different structure. It seems reasonable, but this has hampered to perform sound comparative evaluations and cross-lingual learning experiments. It is reported that statistical parser output in one language cannot be easily compared or transferred to another when using two training data which labelled with different annotation schemes (McDonald et al, 2011; Søgaard, 2011). Mcdonald et al. (2013) reported improved results on cross-lingual transfer parsing using 10 uniformly annotated treebanks.

The Universal Dependencies (UD) seeks to develop cross-linguistically consistent treebank annotation guidelines and apply them to many languages to create treebank annotations, aiming to capture similarities as well as idiosyncrasies among typologically different languages, and released guideline to assist with the creation of new UD treebanks, or mapping and conversions of existing treebanks to a new universal scheme. The UD scheme is built on the Google Universal part-of-speech (POS) tagset (Petrov et al., 2012), the interset interlingua of morphosyntactic features (Zeman, 2008), and Stanford Dependencies(Tsarfaty, 2013; de Marneffe et al., 2014). In addition to the abstract annotation scheme, UD defines also a treebank storage format, CoNLL-U. The UD scheme accounts for varying linguistic differences across language by providing the option of defining language–specific label sub-types when the prescribed list of labels do not adequately cover all linguistic features of a given language. Nivre (2015) explains the motivation behind the project. Since then, a large number of additional treebanks have been either built or converted from existing treebanks to form new UD treebanks. To date, there are 54 treebanks representing 40 languages listed in the UD project.

We have mapped the Uyghur dependency Treebank (UyDT) (S.Mamitimin et al., 2013; M.Aili et al., 2016) to the UD scheme (Version 1) for purposes of cross-lingual studies and parser improvement. The UyDT is a corpus of Uyghur sentences that have been annotated manually. This paper summarizes the conversion and mapping of the UyDT to Uyghur Universal Dependency Treebank (UyUD), as part of the Universal Dependencies (UD) Project.

Proceedings of WLSI/OIAF4HLT,
pages 44–50, Osaka, Japan, December 12 2016.

2 Brief introduction for UyDT

Uyghur is a Ural-Altaic language, has rich and complex morphological structure. As a typical agglutinative language, Uyghur displays rather different characteristics compared to those more well-studied languages in the parsing literature. On the syntactic side, Uyghur has SOV constituent order, and considered a free-constituent order language. Uyghur is also a pro-drop language, as the subject can be elided if necessary, and recovered from the agreement markers on the verb.

We aim at building a dependency treebank to provide basic resources for future NLP researches. Morphological structure plays an important role in finding syntactic relations between words in Uyghur sentences. So all texts are morphologically analysed by Uyghur Morphological Analyser (UMA) software (M. Aili et al., 2012), There are 13 basic POS tags as shown in Table 1.

No	tags	POS	No	tags	POS
1	N	Noun	7	I	Imitative
2	A	Adjective	8	C	Conjunction
3	M	Numeral	9	T	Particle
4	Q	Quantifier	10	E	Exclamation
5	D	Adverb	11	V	Verb
6	P	Pronoun	12	R	Postposition
			13	Y	Punctuation

Table 1. Basic Post Tags in Uyghur Languages

There are 23 dependency relations scheme in UyDT as general as possible which are listed in Table 2.

No.	Label	Relations	*No.*	Label	Relations
1	ABL	Ablative Adjunct	13	OBJ	Object
2	ATT	Attributive Modifiers	14	POSS	Possessor
3	ADV	Adverbial Modifier	15	POST	Postpositions
4	APPOS	Apposition	16	QUOT	Quotation
5	AUX	Auxiliary Verb	17	ROOT	ROOT of Sentence
6	CLAS	Classifier	18	PRED	predicate
7	COLL	Collocation	19	SUBJ	Subject
8	CONJ	Conjunction	20	CL	Clause
9	COORD	Coordination	21	IND	Independent component
10	DAT	Dative Adjunct	22	COP	Copula
11	INST	Instrumental Adjuncts	23	COMP	Comparison
12	LOC	Locative Adjunct			

Table 2. Dependency relation tags in Uyghur Dependency Treebank

3 mapping

3.1 mapping POS-tagset

The UD part-of-speech (POS) tagset is an extension of The Google Universal POS tagset (Petrov et al., 2012) and contains 17 POS tags, whereas, in UyDT, there are only 13 POS tags. Fortunately, we could map most of them to Universal POS tags (e.g. N→ Noun, A→ADJ).

However, only 10 POS tags in UyDT are mapped one by one to UD POS tags, six of the UD POS tags are not used, two tags in Uyghur POS tags are mapped to a same UD POS tag , as : (1) we didn't identify auxiliary verbs in Uyghur which is actually a verb and called auxiliary verb only when combining with other substantive word and indicating a grammatical meaning ; (2) In UyDT POS tagset, pronoun is also tagged as noun, as a result, PROPN in UD POS tags is also not used as well; (3) there are some discussion about DET, as there is not a tag called DET in Uyghur POS tagset, but some words have the meaning in a specific situation, which are numbers most of time. (4) Other three tags (SCONJ , SYM and X) are not used in UyDT. (5) According to the description of INTJ, two tags in

UyDT (exclamation and imitative) matched with it. We provide a mapping from the Uyghur POS tagset to the UD tagset in Table 3.

UD	UyDT tag	UyDT POS	UD	UyDT tag	UyDT POS
ADJ	A	Adjective	NUM	M	Numeral
ADV	D	Adverb	PART	T	Particle
ADP	R	Postposition	PRON	P	Pronoun
*	Q	Quantifier	PUNCT	Y	Punctuation
CONJ	C	Conjunction	VERB	V	Verb
INTJ	*E* *I*	*Exclamation* *Imitative*	NOUN	N	Noun
PROPN	*		X	*	
AUX	*		SYM	*	
SCONJ	*		DET	*	

Table 3: Mapping of the UyDT's POS tagset to the UD's POS tagset

3.2 mapping relations

UD defines a set of 40 broadly applicable dependency relations, further allowing language – specific subtypes of these to be defined to meet the needs of specific resources. However, there are only 23 types of dependent relations in UyDT. The conversion from UyDT dependency annotation to UD required not only relabelling types, but also changes to the tree structure, obviously, it isn't a straightforward mappings. We use three steps to finish the conversing: rule based automatic label mappings; structural changes; manual checking. The details are as follows:

3.2.1 rule based automatic label mapping

Most of the dependency relations which defined in UyDT are included in the UD, but isn't one by one mapping. After comparing the Uyghur treebank relation description with UD description, we mapped Uyghur DT dependent relations to UD as following table. The relation 'ATT', for instance, could map to 'acl, amod, det, nummod', which of them should be chosen is another problem. To tackle with this problem, we settled priority and some limited rules on them according to our corpus features to choose one of them.

Uyghur	Universal	Uyghur	Universal
ATT	acl, amod, det, nummod	ADV	advcl, advmod
CL	advcl, parataxis	APPOS	appos
AUX	aux	POST	case
CONJ	cc	QUOT	ccomp
COLL	compound, mwe, list, name, nummod, goeswith	COORD	conj
COP	cop, neg	PRED	nsubj
IND	discourse, parataxis, vocative	OBJ	dobj, nmod:cau
LOC	nmod	DAT	nmod
COMP	nmod:comp	POSS	nmod:poss, nmod:part, nmod:poss
LOC	nmod:tmod	SUBJ	nsubj
ROOT	punct		

Table 3 Mapping of the UyDT dependent relation to UD dependent relation

For example, the dependent relation ‘OBJ’ in UyDT could map to ‘dobj, dobj:cau, nmod:cau’ in UD, considering that the rate of using causative word is less than using non-causative word, we decided map all the dependent relation ‘OBJ’ to ‘dobj’; the dependent relation ‘ATT’ in UyDT could map to ‘acl, amod, det, nummod’. After adding some limitation on the dependent relation ‘ATT’, such as when the word is tagged ‘NOUN’, map it to ‘amod’, when it is tagged ‘NUM’ map it ‘amod’, and tagged ‘PRON’ map it to ‘det’. After rule based mapping, most of the dependent relations are transformed correctly, certainly including some wrong labels as well. Then, we manually checked and corrected them.

3.2.2 structural changes

The UD syntactic annotation is based on the universal Stanford Dependencies (SD) scheme (de Marneffe et al., 2014). One of the key properties of these schemes is that they emphasizes direct relation between content words, treating function words as dependents of content words rather than as their heads. However, it is not all the case in UyDT. Some function words such as copula or auxiliary words were head of the predicative, for when a copula or auxiliary attaching a word, it would indicate a grammatical meaning as well as get certain morphological forms. For example: ‘*u hetni yezip* ***boldi*** (he had written the letter); *yezip* ***bolghan*** *hetni oqudi* (he read the letter which had been written)’. In these examples, the word with bold font, generated from one stem *bol*, has different morphological form in each sentences to combine these words around it. Though it is auxiliary verb, produce relation ‘aux’ and marked as the head of the relation in UyDT. It contrasts with UD and needs to make some structural changes. We done this changes with manually, for structural changes were not easily automated. The following structural changes were made manually:

- aux & cop
  In the UyDT, the auxiliary and copula are treated similarly to a verb, and can function as the root of a sentence. However, the UD scheme analyses copula constructions differently: the predicate is regarded as the head of the phrase, and the auxiliary or copula is its dependent, as labelled by the ‘aux’ or ‘cop’. See Figure.1（a) and (b) for comparison.

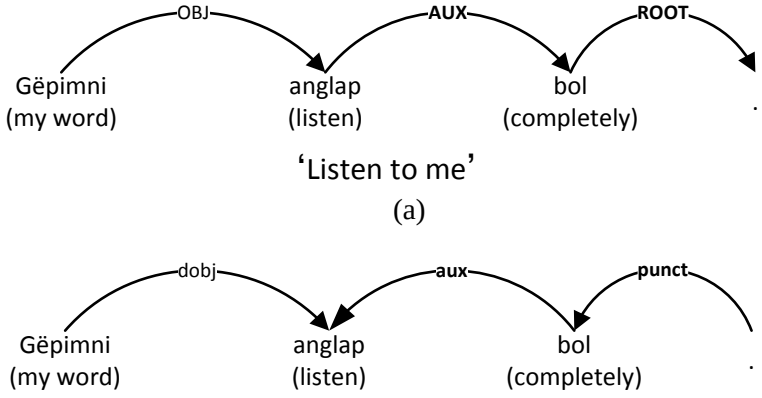
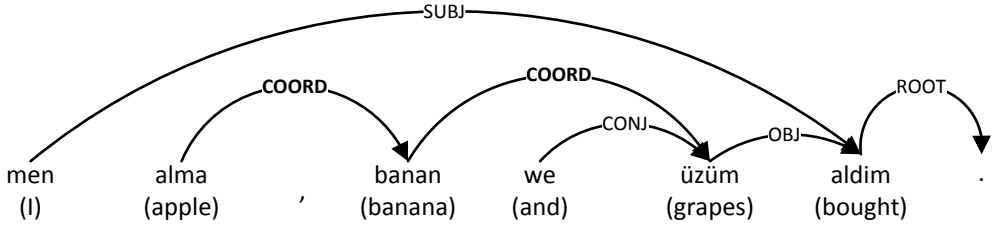


Figure 1: UD aux analysis

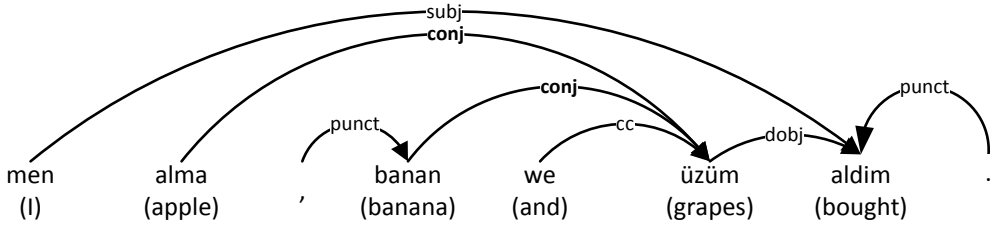
- punct
  In the UyDT, the punctuations which appeared in the sentence was not considered in dependent relation, instead, the last punctuation which appeared the end of a sentence was regarded as the head of the sentence and labelled as ‘ROOT’. However, the UD defines a punctuation depend on content word which it always attached to with the relation of ‘punct’ and can never have dependents. It is need to change the relation structure and the label of the relation ‘ROOT’ in UyDT. (Figure 1)
- conj & cc
  Significant changes were made to the analysis of coordination. In the UyDT， defined words which formed coordinate relations depended from begin to end relatedly and the last one was

the head of them with the label of ‘COORD’ . Meanwhile, the conjunction was depend on the coordinate word which it attached to with the label as ‘CONJ’ (Figure 2 (a)). The UD annotation scheme, on the other hand, uses right-adjunctions, where the first coordinate is the head of them, and the rest of phrase is adjoined to the right. We diverge from UD specification by marking the last conjunct as the head of the relation. All the other conjuncts depend on the last via labelling subsequent coordinates as ‘conj’ (Figure 2 (b))



‘I bought apples, bananas and grapes.’

(a)



‘I bought apples, bananas and grapes.’

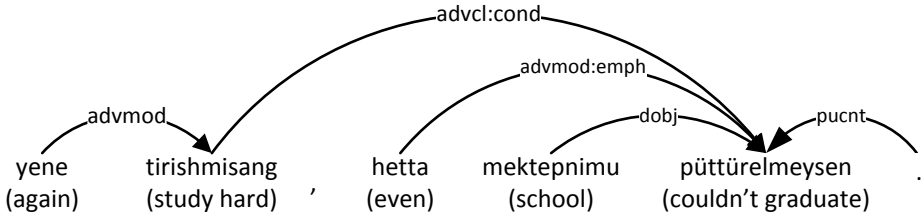
(b)

Figure 2: Coordination structure in the UD

3.2.3 Uyghur-specific relations

The UD scheme provides scope to include language-specific subtype labels. The label naming format is *unversal:extention*, which ensures that the core UD relation remains identifiable, making it possible to revert to this coarse label for cross-lingual analysis. During the conversion of the UyDT, we defined some labels required to represent Uyghur syntax more concisely. These labels are discussed below:

- advmod:emph
  Some adverbial modifiers in Uyghur has served as the emphasizer or intensifier. We use the subtype label ‘advmod:emph’ in cases where modifiers emphasize or intensify their heads. It is also used in the Turkish, Ancient Greek, Arabic, Czech, Latin, Portuguese and Tamil scheme as well. (Figure 3)



If you didn’t study hard futher, you even couldn’t graduate

Figure 3: UD advmod and advcl analysis

- advcl:cond
  It is used for conditional clauses. It is also used in Turkish scheme. (Figure 3)
- aux:q

In Uyghur, a question sentence is built by adding one of question particle to predicate (auxiliary verb or copula). We use ‘aux:q’ for all uses of the question particle. It also used in Hebrew, Turkish. (see Figure 4)

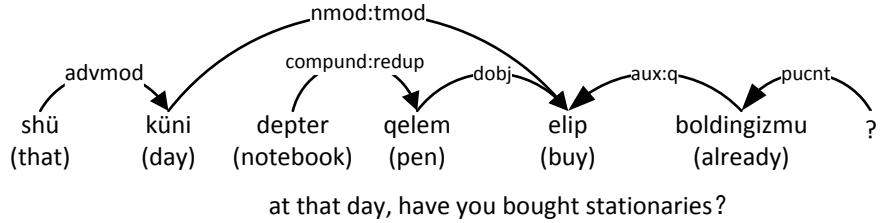


Figure 4: UD aux, compound and nmod analysis

- compound:redup
  Reduplication is a common process especially for adverbs, adjectives, nouns in Uyghur. Reduplication typically involves two identical words, but some morpho- phonological alternations are possible. The forms of the reduplicate words in Uyghur are various, this subtype of compound covers a range of reduplicated forms in Uyghur. It is also used in Turkish as well. An example is given in Figure 4.
- dobj:cau & nmod:cau
  We mark direct objects of causative verbs with ‘dobj:cau’, since the interpretation is different in comparison to a direct object of a non-causative verb. In general, if the verb is intransitive, direct object indicates the “causee”, the subject of the content verb, or the entity that performs the action. If the verb is transitive, the direct object is the entity that is acted upon as in the non-causative case use the subtype ‘nmod:cau’ . They are also used in Turkish as well.
- nmod:tmod
  Temporal modifiers specifying time, in nominal form, are labelled as ‘nmod:tmod’. English, Chinese, Danish, Russian etc. also uses this subtype label. See the Figure 4 for example.
- nmod:poss
  This subtype is used in possessive constructions, typically, the head of the construction is a possessive noun phrase, and the dependent is in genitive case. Danish, English, French, German, Kazakh etc. also use the subtype. An example is giben in Figure 5.
- nmod:comp
  This subtype of ‘nmod’ is used for marking comparative modifier of an adjective or adverb. The specific feature of it is a nominal word or phrase which attached ablative case suffix and an adjective or adverb. This subtype is also used in Turkish as well. See the Figure 5 for example.

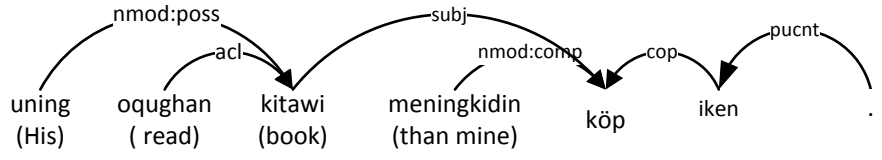


Figure 5: UD nmod relation analysis

- nmod:part
  This subtype of nmod is used for marking the part-whole relations. This structure is similar to ‘nmod:poss’ in most cases, but the range structures expressing “part of” is diverse, and distinction is often be useful.

4 summary and future work

In this paper, we have summarized the conversion of the Uyghur Dependency Treebank (UyDT) to UD format. We have described in detail the mapping and conversion process, including structural

changes required, for the release of the UyDT as part of the Universal Dependencies project. We have also discussed linguistic analyses and motivations for choosing of Uyghur language-specific label types.

Acknowledgments

This work was funded by the Natural Science Foundation of China (Grant No. 61262061) and supported by Science & Technology Foundation of Xinjiang(Grant No. 201423120).

We are extremely thankful to the mathematical and physical department in Charles University and in particular to Dan Zeman for his advice on the Uyghur conversion effort.

Reference

Aili, M., Xialifu, A., Maihefureti, & Maimaitimin, S. (2016). Building Uyghur Dependency Treebank: Design Principles, Annotation Schema and Tools. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Vol. 9442, pp. 124–136).

de Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., & Manning, C. D. (2014). Universal Stanford Dependencies: A cross-linguistic typology. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), 4585–4592.

Džeroski, S., Erjavec, T., & Ledinek, N. (2006). Towards a Slovene dependency treebank. Proc. of the Fifth Intern. …, (May), 1388–1391.

Hajičová, E. (1998). Prague Dependency Treebank: From Analytic to Tectogrammatical Annotation. In Proceedings of the First Workshop on Text, Speech, Dialogue (pp. 45–50).

Haverinen, K., Viljanen, T., Laippala, V., Kohonen, S., Ginter, F., & Salakoski, T. (2010). Treebanking Finnish. In In proc. of The Ninth International Workshop on Treebanks and Linguistic Treories (TLT-9) (pp. 79–90).

Kromann, M. T. (2003). The Danish Dependency Treebank and the DTAG Treebank Tool. In Proceedings of the Second Workshop on Treebanks and Linguistic Theories (pp. 217–220).

Lynn, T., & Foster, J. (2016). Universal dependencies for Irish. In Celtic Language Technology Workshop (pp. 79–92).

Mairehaba·, A., Jiang, W., Wang, Z., Tuergen·, Y., & Liu, Q. (2012). Directed Graph Model of Uyghur Morphological Analysis. Journal of Software, 23(12), 3115–3129.

Mamitimin, S., Ibrahim, T., & Eli, M. (2013). The Annotation Scheme for Uyghur Dependency Treebank. 2013 International Conference on Asian Language Processing, 185–188.

Mcdonald, R., Nivre, J., Quirmbach-brundage, Y., Goldberg, Y., Das, D., Ganchev, K., … Lee, J. (2013). Universal Dependency Annotation for Multilingual Parsing. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, 92–97.

McDonald, R., Petrov, S., & Hall, K. (2011). Multi-source transfer of delexicalized dependency parsers. Proceedings of the Conference on Empirical Methods in Natural Language Processing, (2007), 62–72.

Oflazer, K. (2003). BUILDING A TURKISH TREEBANK, 1–17.

Petrov, S., Das, D., & Mcdonald, R. (2012). A Universal Part-of-Speech Tagset. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC '12), 2089–2096.

Pyysalo, S., Kanerva, J., Missilä, A., Laippala, V., & Ginter, F. (2015). Universal Dependencies for Finnish. Nordic Conference of Computational Linguistics NODALIDA 2015, (Nodalida), 163.

Søgaard, A. (2011). Data Point Selection for Cross-Language Adaptation of Dependency Parsers. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT '11): Short Papers, 682–686.

Tsarfaty, R. (2013). A Unified Morpho-Syntactic Scheme of Stanford Dependencies. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 578–584.

Zeman, D. (2008). Reusable Tagset Conversion Using Tagset Drivers. Lrec, 213–218.