A FST Description of Noun and Verb Morphology of Azarbaijani Turkish

Razieh Ehsani **Berke Özenç** **Ercan Solak**
Işık University, Istanbul, Turkey
name.surname@isikun.edu.tr

Abstract

We give a FST description of nominal and finite verb morphology of Azarbaijani Turkish. We use a hybrid approach where nominal inflection is expressed as a slot-based paradigm and major parts of verb inflection are expressed as optional paths on the FST. We collapse adjective and noun categories in a single nominal category as they behave similarly as far as their paradigms are concerned. Thus, we defer a more precise identification of POS to further down the NLP pipeline.

1 Introduction

Azarbaijani Turkish (AT) is a Turkic language spoken by more than 30 million people mainly in Iran and Azarbaijan. AT is an agglutinative language with rich and regular inflectional and derivational morphologies. As in all Turkic languages, vowel harmony and consonant changes at the morpheme boundaries are conditioned by the phonological context. Word order is relatively free and the syntactic relations are indicated by case markings and in spoken form also by stress.

The alphabet has 23 consonants, adding ç, ş, ğ to the consonants of the Latin alphabet and removing w from it. It has 9 vowels, adding ı, ü, ö and ə to the Latin alphabet. The Table 1 gives the frontness, roundness and height of vowels.

	Front		Back	
	Flat	Round	Flat	Round
Close	i	ü	ı	u
Mid	e	ö		o
Open	ə		a	

Table 1: Vowels and their properties

There are a few morphological analyzers for Turkish, (Oflazer, 1994), (Çöltekin, 2010), (Şahin et al., 2013) and Turkic languages like Turkmen (Tantug et al., 2006), Kazakh (Kessikbayeva and Cicekli, 2016), Uighur (Orhun et al., 2009). Although, AT is close to the Turkish spoken in Turkey, there are enough non-trivial differences both in morphotactics and phonology to prevent the direct use of analyzers implemented for Turkish. To the best of our knowledge, this work is the first FST implementation of AT noun and finite verb inflections.

In our implementation, we used Helsinki FST, (Lindén et al., 2011). We provide the details of the full FST as supplementary materials. The present description involves nominal and verb inflections in isolation. We are extending this initial effort to the rest of the AT morphology and we will make the full implementation publicly available as a web service.

2 Approach

For each morpheme, we represent its abstract form either as a key-value pair for slot-based paradigms or just as a key for other paradigms. For example, <Case:Abl> denotes an abstract morpheme for ablative case. The first level of morphology yields the archmorphemes prior to the phonological transformations in the second level. Thus, in FST description, a transition expressed as <Case:Abl>:-NAn yields the archmorpheme -NAn where the archiphoneme 'N' stands for a choice of 'n' or 'd' and 'A' stands for 'a' or ə.

3 Nominal inflection

The nominal inflection in AT has a fixed order of suffixes as
Nominal stem + Number + Possessor + Family + Case.

Nominal stem may be either a simple nominal

Proceedings of the 13th International Conference on Finite State Methods and Natural Language Processing, pages 64–68,
Umeå, Sweden, 4–6 September 2017.
https://doi.org/10.18653/v1/W17-4008

root or a complex form that has already undergone a series of derivations.

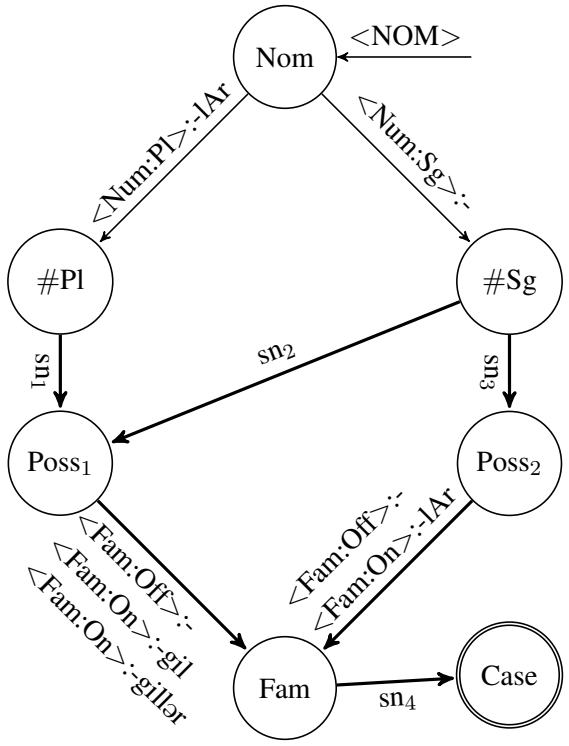

Figure 1: Slot-based nominal inflection in AT

The Number slot can have Singular or Plural values where Singular has zero surface form. There are 7 values for the Possessor key including the value None for no possessor. The Case slot may have values Nominal, Accusative, Dative, Locative, Ablative, Genitive and Instrumental.

The Family slot is binary valued as On/Off. We could have represented it as an optional morpheme without a slot. However, as we wanted to represent all of nominal inflection as slot based, we used the binary value trick to represent its presence.

When On, the surface realizations of the Family slot are one of -gil, -gillər or -lAr. Note that the third form -lAr is the same as the surface realization of Plural morpheme. The collision of surface realizations is the reason for the occurrence of such seemingly irregular forms as

(1)	xala-m-lar	xala-lar-ım
	aunt-P1SG-PLU	aunt-PLU-P1SG

The apparent reordering of the possessor and plural suffixes in (1) seems to contradict the strictly ordered paradigm of nominal inflection. However, the semantics of the two forms (1) indicate the presence of two distinct morphemes with the same surface forms. Indeed, while "xala-lar-ım" means "my aunts", "xala-m-lar" means "the family of my aunt" or "my aunt and her group." The analysis with the family slot becomes more apparent when we use the other family suffix -gil

(2)	xala-m-gil	*xala-gil-im
	aunt-P1SG-FAM	aunt-FAM-P1SG

where the family suffix is forced to come after possessor.

The FST for the nominal paradigm is given in Figure 1.

The possessor morphemes sn_1, sn_2 and sn_3 in Figure 1 are listed in Table 2. Note that some morphemes such as <poss:2p> have multiple equivalent surface forms.

Abstract morpheme	**First level output**		
	sn_1	**sn_2**	**sn_3**
<Poss:None>	-	-	-
<Poss:1s>	-Im	-(I)m	-(I)m
<Poss:2s>	-In	-(I)n	-(I)n
<Poss:3s>	-I(n)	-(s)I(n)	
<Poss:1p>	-ImIz	-(I)mIz	
<Poss:2p>	-InIz	-(I)nIz	
	-Iz	-(I)z	
<poss:3p>	-I(n)	-lArI(n)	

Table 2: Possessor person paradigms for Figure 1. Empty cells are undefined. Apart from epenthesis, the only difference between sn_1 and sn_2 is <Poss:3p>.

Since our nominal inflection paradigm is slot-based, the only accepting state is the Case. The case symbols for sn_4 are of the form <Case:v>:a where (v,a) pair is one of (Nom,-), (Dat,-(y)A), (Loc, -dA), (Acc, -(n)I), (Abl, -NAn), (Gen, -(n)In) and (Ins, -(y)InAn).

4 Verbal paradigm

The inflection of finite verbs also has a fixed order with few exceptions mainly due to the copulas. In its general form, the verb paradigm is
Verb stem + Voice + Ability + Polarity + Probability + Tense-aspect-mood + Person.

The whole paradigm is quite complex. In order handle the complexity, we divided the paradigm into smaller sub-paradigms with clear interfaces among them.

4.1 Voice

There are 5 voices, Active, Reflexive (Rflx), Reciprocal (Rcpr), Causative (Caus) and Passive (Pasv). Apart from Active voice, each has a variety of surface realizations which are specified lexically or phonologically. A verb stem might have multiple voices under co-occurrence and ordering restrictions. The ordering of the voices are

(Reciprocal or Reflexive) + Causatives + Passive.

Theoretically, the Causative voice marker can be repeated freely to denote an arbitrarily long chain of causation. However, in practice, only up to three Causative markers are used.

The FST for the voice paradigm is shown in Figure 2. The Causative and Passive morpheme symbols sv_1, sv_2 and sv_3 in Figure 2 are listed in Table 3.

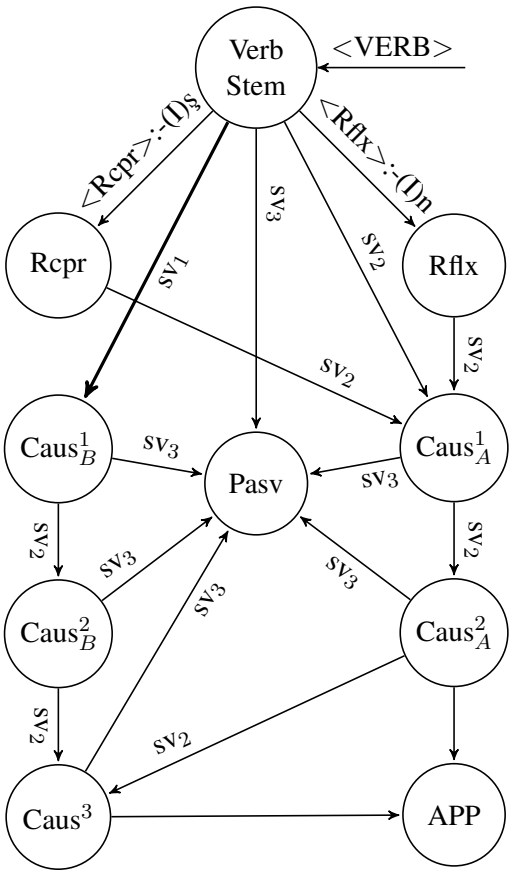

Figure 2: Voice paradigm

The APP state in Figure 2 represents the state connecting the Voice FST to the part of the verb inflection paradigm which deals with Ability, Polarity and Probability given in Figure 3. We treat the transitions within the voice paradigm as optional so all the states have an edge to APP state. In order not to clutter the diagram, we showed only two of these, the ones from Caus^3 and Caus^2_A. There are two distinct paths for multiple causatives, the path Caus^1_A-Caus^2_A-Caus^3 and Caus^1_B-Caus^2_B-Caus^3. These are conditioned by the phonotactics of the different causative suffixes. However, it is difficult to come up with rules governing the choice of a particular causative suffix. Less common suffixes are best represented as being conditioned by the verb stem lexicon. Note that there are 4 allomorphs (rows sv_1 and sv_2 in Table 3) for causative and 2 (row sv_3) for passive when they are attached immediately after the verb stem. In our implementation, we created 15 distinct initial states to handle their possible combinations (including no Passive and no Causative) and we marked the correct transition within the verb lexicon. We refrained from using flag diacritics at the expense of increasing the number of states.

Symbol	Morpheme
sv_1	<Caus>:-Irt
	<Caus>:-Art
	<Caus>:-(I)t
sv_2	<Caus>:-dIr
sv_3	<Pasv>:-(I)l
	<Pasv>:-(I)n

Table 3: Voice suffixes

4.2 Ability, Polarity, Probability

The interaction among Possibility, Polarity and Probability is given in the FST in Figure 3. The multiple paths for polarities are needed for the different Person paradigms of negative Aortive tense. The symbol <Abil> denotes the Ability abstract morpheme. Similarly, <Prbl> denotes Probability morpheme. <Pol:Pos> and <Pol:Neg> denote positive and negative polarity morphemes, respectively. <Pol:Pos> has zero surface realization.

In AT, there are two morphemes with the same canonical surface forms to express Possibility and Probability. These respectively correspond to 'be able to' and 'might' in English. Similarity of surface forms precludes their adjacency. So, it is impossible to morphologically compose 'he might be able to come' in AT. Instead, we end up with an ambiguous construction that expresses either possibility or probability. However, when the negative polarity suffix -mA is in between, we can say

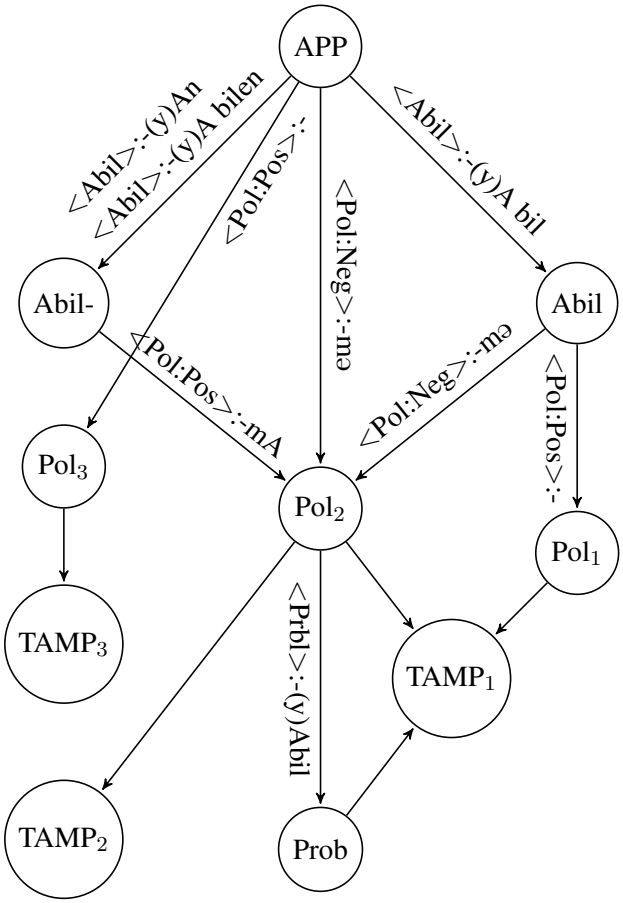

Figure 3: Ability-Polarity-Probability

(3) gəl-ə bilən-mə-yebil-ər
come-ABIL NEG-PROB-AOR-3SG
he might not be able to come.

The standard orthography of AT has a split of the word before the surface form 'bilən' of Ability morpheme. Furthermore, there are two more allomorphs -(y)An and -(y)A bilən of Ability morpheme when it is followed by negative polarity.

The states labeled as $TAMP_1$, $TAMP_2$ and $TAMP_3$ in Figure 3 are the states connecting the FST describing Ability, Polarity and Probability to the FST describing Tense, Aspect, Modality and Person (TAMP) which is partially given in Figure 4. Note that the transitions to TAMP states do not output anything.

4.3 Tense, Aspect, Modality and Person

The last paradigm of finite verb inflection has the basic order
Tense + Copula + Person + Condition + Question.

Only Tense and Person are obligatory and the rest are optional. The order of Copula and Person changes in some cases.

In AT, often a single morpheme expresses a combination of tense, aspect and modality (TAM). There are 11 tenses. In terms of their interactions with their surrounding context, we grouped them into 9 distinct groups. There are 4 copula morphemes corresponding to Aortive, Narrative, Past and Conditional.

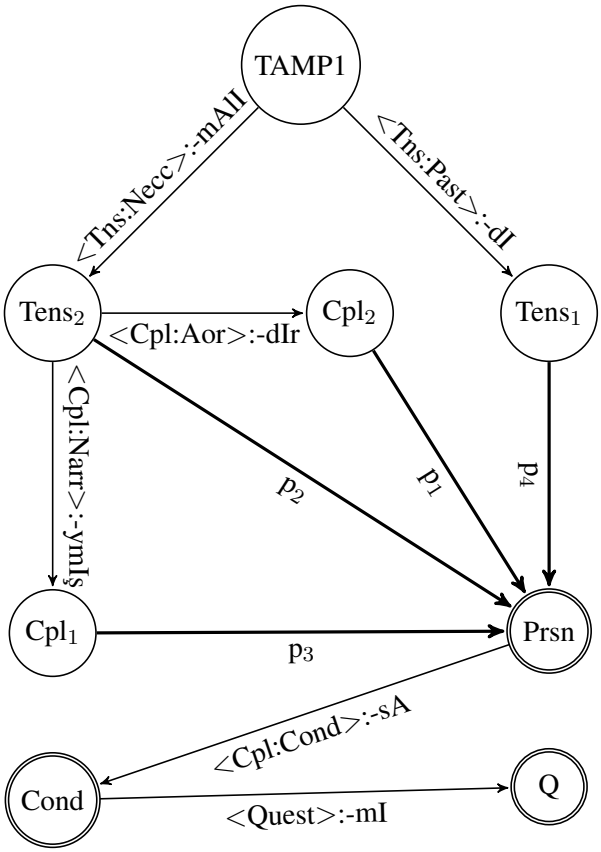

Figure 4: Tense-aspect-mood and person paradigm for Past and Necessity tenses.

The FST for the full TAM-Person paradigm is too complicated to present in the body of the present paper. We provide in Figure 4 only a part of it to show different paths for copula and person paradigms. We provide the full paradigm in the supplementary materials.

Person	**p**$_1$	**p**$_2$	**p**$_3$	**p**$_4$
<Prsn:1s>		-yAm	-Am	-m
<Prsn:2s>		-sAn	-sAn	-n
<Prsn:3s>	-		-	-
<Prsn:1p>		-yIQ	ıQ	-q
<Prsn:2p>		-sInIz	-sInIz	-nIz
<Prsn:3p>	-lAr		-lAr	-lAr

Table 4: Person morphemes in Figure 4.

In Figure 4, <Tns:Past> and <Tns:Necc> represent abstract morphemes for Past and Necessity tenses, respectively. <Cpl:Narr> and <Cpl:Aor> represent the copula morphemes for Narrative and Aortive.

The possible person suffixes p_1-p_4 are listed in Table 4. Note that <Prsn:3s> has zero surface realization. The empty cells in columns p_1 and p_2 show that following Necessity tense, the Aortive Copula is possible only for singular or plural third person.

5 Phonology

In AT, the rules governing phonemic changes depend mostly on the phonological context, enabling an almost complete decoupling of morphotactics and phonology in the implementation. In our implementation, however, we embedded some of the phonological rules in the morphotactics FST whenever the phonological context is known.

There are two major categories of phonological rules in AT. The first one deals with the insertion of epenthetic letters (y), (n) and (s). In our implementation, we start with epenthetic letters inserted by default and drop them when needed. The second major category deals with the mapping of archiphonemes, A, I, N, Q and K to their surface forms. We list the rules below in the order they are applied.

1. Epenthetic (y) and (s) drop when they follow a consonant.
2. Epenthetic (n) drops when it is the last letter in a word.
3. The negative suffix -mA drops its A before a vowel or (y).
4. The archiphoneme N maps to n when it follows n or m, it maps to d otherwise.
5. The archiphoneme A is mapped to surface phonemes to satisfy back-front harmony. It maps to a when it is the first vowel after a back vowel and it maps to ə when following a front vowel.
6. The archiphoneme I maps to u, ü, ı or i for round-flat and back-front harmony.
7. The archiphoneme Q maps to the archiphoneme K when it follows a front vowel. K is further mapped according to the next rule.
8. The archiphoneme Q maps to ğ and the archiphoneme K maps to y before a vowel.

6 Conclusion

We provided HFST descriptions of the nominal and finite verb inflections of Azarbaijani Turkish. Our specification of nominal inflection resolves the apparent reordering of possessive and plural morphemes using a new family/group morpheme.

Currently, we are working on working out the rest of the morphology complete with derivation, nominal predicates and a root lexicon. After unifying the isolated parts in a single analyzer, we plan to make its implementation publicly available and also present the analyzer as a web service.

References

[Çöltekin2010] Çağrı Çöltekin. 2010. A freely available morphological analyzer for Turkish. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 820–827.

[Kessikbayeva and Cicekli2016] Gulshat Kessikbayeva and Ilyas Cicekli. 2016. A Rule Based Morphological Analyzer and a Morphological Disambiguator for Kazakh Language. *Linguistics and Literature Studies*, 4(1):96–104.

[Lindén et al.2011] Krister Lindén, Erik Axelson, Sam Hardwick, Miikka Silfverberg, and Tommi Pirinen. 2011. HFST–framework for compiling and applying morphologies. pages 67–85.

[Oflazer1994] Kemal Oflazer. 1994. Two-level description of turkish morphology. *Literary and Linguistic Computing,*, 9(2):137–148.

[Orhun et al.2009] Murat Orhun, A. Cüneyd Tantug, and Esref Adali. 2009. Rule based analysis of the uyghur nouns. *Int. J. of Asian Lang. Proc.*, 19(1):33–44.

[Şahin et al.2013] Muhammet Şahin, Umut Sulubacak, and Gülşen Eryiğit. 2013. Redefinition of turkish morphology using flag diacritics. In *Proceedings of The Tenth Symposium on Natural Language Processing (SNLP-2013)*, Phuket, Thailand, October.

[Tantug et al.2006] A Cüneyd Tantug, Esref Adali, and Kemal Oflazer. 2006. Computer Analysis of the Turkmen Language Morphology. *FinTAL*, 4139:186–193.

A Supplemental Material

We provide the diagrams for the full FST's as supplementary material.

The supplementary materials are available at `http://www2.isikun.edu.tr/personel/ercan.solak/MorAz/index.html`