A Syntactically Expressive Morphological Analyzer for Turkish

Adnan Öztürel
Google Research
ozturel@google.com

Tolga Kayadelen
Google Research
tkayadelen@google.com

Işın Demirşahin
Google Research
isin@google.com

Abstract

We present a broad coverage model of Turkish morphology and an open-source morphological analyzer that implements it. The model captures intricacies of Turkish morphology-syntax interface, thus could be used as a baseline that guides language model development. It introduces a novel fine part-of-speech tagset, a fine-grained affix inventory and represents morphotactics without zero-derivations. The morphological analyzer is freely available. It consists of modular reusable components of human-annotated gold standard lexicons, implements Turkish morphotactics as finite-state transducers using OpenFst and morphophonemic processes as Thrax grammars.[1]

1 Introduction

The agglutinative morphology of Turkish is complex, due to rich inflectional and derivational morphotactics, a considerably large affix inventory, and morphophonemic processes with potential irregularities. Therefore, morphology processing is an integral part of Turkish NLP in devising sublexical representations to serve the needs of language model development (Oflazer et al., 2003; Çakıcı, 2005; Sulubacak et al., 2016).

From a theoretical standpoint, Bozşahin (2002) claims that transparent integration of morphology to syntactic processing is essential in order to overcome phrasal scope conflicts. They propose that morphology-syntax integration can be attained in architectural level using: (i) a lexemic grammar where morphological parsing is the precursor of syntactic analysis to resolve sublexical hypothesis space for syntax to operate on lexemic constituents, or (ii) a morphemic grammar with lexical items of root forms and affixes that has adequate lexical categories to capture correct semantic bracketing, for a transparent morphology-syntax interface. They illustrate latter approach on a linear fragment of Turkish inflectional paradigms using a lexicalized grammar formalism.

The former approach is studied mainly over two-level models (Koskenniemi, 1984). Oflazer (1994) presents the first two-level description of Turkish morphology, Sak et al. (2009) adapts this definition to build a stochastic finite-state transducer (FST) that is trained on 200 million words and Şahin et al. (2013) utilize flag diacritics in limiting illicit morphological parses. Considering the restricted availability of these morphological analyzers, open-source alternatives have been proposed by Akın and Akın (2007) and Çöltekin (2010, 2014).

In this paper we present a morphology model for Turkish that improves the above-mentioned models in a number of ways. Our model captures all syntactic processes that are handled by morphology at the word level over a sufficiently elaborate representation. It uses a gold standard human-annotated lexicon which, to our knowledge, is the first in the literature. We introduce a fine part-of-speech tagset which provides finer control in modeling morphotactics for lexical categories, and represent productive derivational morphology in a level of comprehensive scrutiny that none of the previous models do. Finally, we present novel methods to represent named entities in morphological analysis, eliminate zero-derivations from morphotactics and a linguistically sound approach to handle some intricacies around case morphology.

The model is implemented as an FST, it is open-source, thus extensible. It can be used in building lexemic syntactic processors that depend on morphological analysis, and also in morphemic grammar development and treebank induction.

[1] https://github.com/google-research/turkish-morphology

Proceedings of the 14th International Conference on Finite-State Methods and Natural Language Processing, pages 65–75
Dresden, Germany, September 23-25, 2019. ©2019 Association for Computational Linguistics

Input:	affıyla
Intermediate:	af”+SH+YlA
Output:	(af[NN]+[PersonNumber=A3sg]+SH[Possessive=P3sg] +YlA[Case=Ins])+[Proper=False]

Figure 1: Levels of analysis for the word *affıyla* ‘*with their excemption*’. For illustrative purposes ambiguous interpretations on both intermediate and output tape is omitted and only a single parse is presented.

2 Levels of Analysis

Morphological analysis is composed of morphophonemic and morphotactic analysis layers. As illustrated in Fig. 1 the morphophonemic layer acts as the first level of analysis. It resolves phonetic processes that work at the morphology level by mapping input surface forms to an intermediate representation (see Section 3). The intermediate representation consists of an annotation of the morphophonemic irregularities of the root followed by the meta-morphemes that correspond to the affixes that are realized in the surface form.[2] The morphotactic layer is composed of the lexicon of root forms (see Section 4), affix inventory, and a word-internal grammar that defines affixation paths for each lexical category (see Section 5). It maps the intermediate representation into a morphological parse, which represents the sublexical segmentation and marks the root form with its lexical category, and inflectional and derivational affixes with their functional feature tags.

3 Morphophonemics

The morphophonemic layer is implemented as 9 Thrax grammars (Roark et al., 2012) which are formed of regular expressions and word-internal context-dependent rewrite rules that are compiled into FSTs. Composing the FSTs defined by these grammars yields the morphophonemic model. We handle all known phonological phenomena that play a role in Turkish word formation and that manifest itself in word orthography (Oflazer et al., 1994; Göksel and Kerslake, 2004).

A vowel harmony grammar maps back/front vowels into the meta-phoneme A and high vowels to H given the preceeding vowels (e.g. *evinde* → *ev***H***nd***A**). A vowel change grammar implements the alteration of root final ‘*e*’ to ‘*i*’ when a suffix that starts with ‘*y*’ is affixed (e.g. *diyecek* → *d***e***yecek*). A vowel drop grammar implements elision, i.e. /vowel/ - /∅/ alteration (e.g. *burnu* → *bur***u***nu*).

A consonant voicing grammar handles sonorization and respectively maps root final {‘*t*’, ‘*d*’} into {‘*p*’, ‘*b*’} and {‘*c*’, ‘*ğ*’, ‘*ng*’} into {‘*ç*’, ‘*k*’, ‘*nk*’} if a suffix starting with a vowel is affixed (e.g. *kitabının* → *kita***p~***ının*, or *rengi* → *re***nk***i*). A consonant change grammar maps suffix initial dental consonants {‘*d*’, ‘*t*’} into the meta-phoneme D by referring whether the morpheme to its left ends with {‘*f*’, ‘*s*’, ‘*t*’, ‘*k*’, ‘*ç*’, ‘*ş*’, ‘*h*’, ‘*p*’} (e.g. *evde* → *ev***D***e*, or *uçakta* → *uçak***D***a*). A consonant drop grammar captures elision of affix initial consonants when the morpheme that preceeds the affix ends with a consonant (e.g. *evinin* → *ev***S***i***N***in*). A gemination grammar implements duplication of the root final consonants {‘*b*’, ‘*d*’, ‘*k*’, ‘*l*’, ‘*m*’, ‘*n*’, ‘*s*’, ‘*t*’, ‘*z*’} when a suffix that starts with a vowel is affixed to the root (e.g. *affıyla* → *af***”***ıyla*). A y-insertion grammar implements insertion of root final ‘*y*’ to roots that end with ‘*su*’ when a suffix starting with a dropping consonant or high vowel is affixed to them (e.g. *akarsuyuyla* → *akarsu*ˆ*uyla*).

Finally, a dedicated morpheme segmentation grammar marks morpheme boundaries (e.g. evlerinde → ev**+**ler**+**i**+**nde). Most of these phonological processes (except vowel harmony and some of the consonant voicing/change processes with certain irregularities) are not generalized but only apply to a small set of roots from certain lexical categories. Therefore, they are annotated on root forms (see Section 4.3).

4 Lexicon of Root Forms

Our lexicon consists of 47,202 entries.[3] An entry is a 5-tuple of root form (or word stem), its part-of-speech (PoS), annotation of morphophonemic irregularities, morphosyntactic and semantic

[2] We represent meta-phonemes in capitals (e.g. **H** represents the set of high vowels {‘*u*’, ‘*ü*’, ‘*ı*’, ‘*i*’}), and fully realized phonemes that appear in the surface form in lowercase. **+** is used in the intermediate representation to denote morpheme boundaries. On the output tape inflectional morphemes are marked with **+** delimeter and derivational morphemes are marked with **-**.

[3] The base lexicon can be extended through open-source contributions especially with lexical items of open class categories. See annotation guidelines on `https://github.com/google-research/turkish-morphology/blob/master/analyzer/src/lexicon/README.md`.

Tag	**Root**	**Morphophonemics**	**Features**	**Compound**
NN	af	af”	-	false
NN	milletvekili	milletvekil	-	true
CC	velevki	-	+[ConjunctionType=Sub]	false

Figure 2: Structure of the lexicon.

features, and a boolean denoting whether the root form is a compound (see Fig. 2).

Each lexicon entry was annotated by 3 human annotators, where one of the annotators was the tie-breaker on 2-way annotation. Thus the lexicon is expected to have higher consistency and quality in compared with those that are acquired through semi-automatic extraction and labeling of lexical items over web-based corpora (Çöltekin, 2014) and affix stripping algorithms (Eryiğit and Adalı, 2004), which do not guarantee gold standard annotations due to the ambiguity that morphophonemic processes introduce in the surface form of the affixes.

4.1 Root Form

By *root form* (or *word stem*) we mean the part of a word form that remains when all inflectional and derivational morphemes are stripped. We assume any productive affixation process should be represented in morphotactics and respective affixes should be members of the affix inventory, but not part of the root form. This includes all morphemes that interact with syntactic processes. Morphosyntactic productivity is not a sole indicator of such processes. Affixes that compositonally alter the semantics of the root form should also be a part of the affix inventory. Our morpheme segmentation scheme, which is based on these principles, is presented in Section 5.2.

4.2 Part-of-Speech Tagset

All previous models of Turkish morphology and labelled corpora assume coarse PoS tagsets (Oflazer et al., 2003; Sulubacak et al., 2016). Distinctively, we use a more elaborate subcategorization of coarse lexical types, the fine PoS tagset that is presented in Table 1. The reason to use a fine categorization is two-fold. It provides control in modeling morphotactics so that we can define a custom grammar of affixation for each lexical category which captures the true inflectional and derivational paradigms of the category in order to restrict overgeneration. Second, the morphological parse incorporates a realistic representation of lexical types and thus it is more informative of the actual syntactic structure.

Coarse Tag	**Fine Tag**	**Description**
ADJ	JJ VJ	Adjective Verb in participle form
ADP	IN	Postposition
ADV	CRB RB WRB	Converb Adverb Interrogative adverb
AFFIX	PFX	Prefix
CONJ	CC	Coordinating conjunct
DET	DT PDT WDT	Determiner Prediterminer Wh-determiner
EXS	EX	Existential verb
NOUN	ADD NN NNP VN	Electronic address Common noun Proper noun Verbal noun
NUM	CD	Cardinal number
ONOM	DUP	Onomatopoeic
PRON	PRD PRF PRI PRP PRP$ PRR WP	Demonstrative pronoun Derived pronoun Indefinite pronoun Personal pronoun Possessive pronoun Reflexive pronoun Wh-pronoun
PRT	EP OP RPC RPNEG RPQ	Final particle Coordinative particle Clitic particle Negation particle Question particle
VERB	NOMP VB	Nominal predicate Verb

Table 1: Fine PoS tagset that is used in lexical categorization. As a reference for comparisong we present their mapping to coarse tags, which is aligned with Universal Dependecies (UD) (Petrov et al., 2012; Nivre et al., 2016) except the bold marked Turkish-specific additions. Due to space considerations we do not present the tags '.' (punctuation) and 'X' (catch-all for abbreviations, etc.). For the complete PoS tagset that we use, refer to `https://github.com/google-research/turkish-morphology/blob/master/analyzer/src/lexicon/README.md`.

The tags are categorized into two mutually exclusive sets. Those that are lexical (used in annotating the PoS of roots in the lexicon), and those that arise due to derivational morphology. The second set is {CRB, PRF, VJ, VN}. Fig. 3-a-d presents an example of their use in sentence-level

(a) **Pronominalization** ‘*Ali took (the one) that is with the child*’

Ali	çocuktakini	aldı
Ali	çocuk+DA-ki+NH	al[VB]+DH
Ali[NNP]	(child[NN]+Loc)(**[PRF]**-Pron+Acc)	take[VB]+Past

(b) **Noun Clause** ‘*Ali knows that you stole the money*’

Ali	parayı	senin	çaldığını	biliyor
Ali	para+YH	sen+NHn	çal-DHk+SH+NH	bil+Hyor
Ali[NNP]	money[NN]+Acc	you[PRP$]+Gen	(steal[NN])(**[VN]**-PastNom+P3sg+Acc)	know[VB]+Prog1

(c) **Relative Clause** ‘*Ali knows the money that you stole since three years*’

Ali	senin	çaldığın	parayı
Ali	sen+NHn	çal-DHk+Hn	para+YH
Ali[NNP]	you[PRP$]+Gen	(steal[VB])(**[VJ]**-PastPart+P2sg)	money[NN]+Acc

üç	yıldır	biliyor
üç	yıl-DHr	bil+Hyor
3[CD]	(year[NN])([RB]-Since)	know[VB]+Prog1

(d) **Adverbial Clause** ‘*I went home running*’

Eve	koşarak	gittim
Ev+YA	koş-YArAk	git+DH+m
Home[NN]+Dat	(run[VB])(**[CRB]**-Ger)	go[VB]+Past+V1sg

(e) **Nominal Predicate** ‘*(that is) Ali’s child*’

Ali’nin	çocuğudur
Ali+’+NHn	çocuk+SH+DHr
Ali[NNP]+Apos+Gen	child**[NOMP]**+P3sg+GenCop

Figure 3: Morphological feature and PoS labeling of sentences that illustrate the use of morphologically derived lexical categories and nominal predicates in sentence-level context.

context. Fig. 3-e illustrates an example for the NOMP (nominal predicate) category. It captures cases where non-verbal roots are affixed with copula markers and act as the main predicate of the sentence. Unlike previous models, we differentiate between verbal and non-verbal predicates in terms of PoS labels.

4.3 Morphophonemic Irregularities

Consonant voicing irregularities apply to roots whose final voiceless consonant fails to get voiced despite attachment of an affix that starts with a vowel. It only applies to sounds that are [-voiced][+plosive]. We annotate final voiceless plosives { ‘*k*’, ‘*p*’, ‘*t*’, ‘*ç*’} on roots that do not follow this process with **K**, **~** and *Ç* (e.g. *meş***K**, *tehdit***~**, *gö***Ç**). Likewise, roots that undergo gemination and y-insertion are respectively annotated with **”** and **^** (e.g. *af***”** or *akarsu***^**).

The lateral ‘*l*’ has allophones when it occurs in root final position after back vowels. When an affix beginning with a vowel is attached to roots with palatalized root final ‘*l*’, affix form is resolved as if the vowel in the last syllable of the root is a front vowel. Hence, we respectively annotate back vowels {‘*a*’, ‘*â*’, ‘*o*’, ‘*u*’} that appear in the last syllable of such roots with **{**, **[**, **%**, and **}** (e.g. *ihtim***{***l* or *metrop***%***l*). Similarly, last vowel of the roots that undergo epenthesis and vowel closing are annotated with **?** and **E** (e.g. *buru***?***n* or *y***E**).

In case of code-switching foreign words are used in Turkish sentences and get inflected according to the lexical category that they hold in sentence-level context while root form is preserved on surface. Last syllable of the Turkish pronunciation of these roots are annotated to guide morpophonemics model to resolve surface form of the affixes that attach to them (e.g. *charter****ır***). Abbreviations are handled in the same manner.

4.4 Lexical Features

Besides the morphological features described in Section 5 we represent certain syntactic agreement, semantic and sentence-level segmentation features in morphological parse. These features are lexically conditioned, thus annotated in the root form lexicon. They can be used in feature-engineering for morphological disambiguation, PoS tagging and syntactic parsing. There are 5 such feature categories:

Apostrophe marks optional apostrophes that separate affixes from nominal and nominal predicate roots (e.g. *Ankara’da* ‘*Ankara*+Apostrophe+Loc’).

Temporal is used to mark common nouns and adverbs that denote temporality (e.g. *süre* ‘(for some) *duration*’ or *akşamüzeri* ‘*towards evening*’).

Input: kitaplık
Output: (kitap[NN]+[PersonNumber=A3sg]+[Possessive=Pnon] +[Case=Bare])([NN]-lHk[Derivation=For]
+[PersonNumber=A3sg]+[Possessive=Pnon]+[Case=Bare])+[Proper=False]

Figure 4: Morphological parse of the word *kitaplık* '*bookshelf*'. Composed of two IGs, each enclosed in parantheses. First one consisting of the root *kitap* '*book*' and its inflections and second consisting of the derivational morpheme -lHk (which derives '*bookshelf*' from '*book*') and its inflectional features.

ConjunctionType specifies subcategorization of conjuct roots, denoting whether they are adverbial, coordinating, parallel or subordinating given the sentence and/or discourse-level context (e.g. *ya* '*either*+Parallel' or *ile* '*with*+Coordinating').

DeterminerType marks determiner roots as definite, indefinite, demonstrative or directional (e.g. *çoğu* '*most of*+Indefinite').

ComplementType indicates whether the complement of a postposition is a number, finite verb, or nominal which is marked for a certain case. This feature is inherited from the METU-Sabancı Treebank (MST) (Atalay et al., 2003; Oflazer et al., 2003). Unlike MST, we distinguish postpositions with number and finite verb complements from those that have nominative case marked nominal counterparts (e.g. (gitti 'went') *diye*+FiniteComplement, or (yatırımcı 'investor') *için*+NominativeComplement).

4.5 Compound Nouns

Certain noun roots end with compounding marker +SH, which is ambiguous with 3rd person possessive inflection morpheme (e.g. *milletvekil(i)* '*member of parliament*+SH'). These roots have irregular nominal inflectional morphotactics. When inflected for 3rd person plural (A3pl), inflectional morpheme +lAr precedes +SH as in Fig. 5. Such noun roots are annotated in the lexicon as shown in Fig. 2 and we define a custom inflectional paradigm for them to capture this behaviour in the morphotactics model.

(a) milletvekil+lAr+SH
milletvekil+ler+i
milletvekilleri
(b) *milletvekil(**i**)+lAr+SH
*milletvekil(**i**)+ler+i
*milletvekilileri

Figure 5: 3rd person singular inflections on compound noun roots.

5 Morphotactics

The morphotactic layer is implemented using the OpenFst library (Allauzen et al., 2007). We define 15 FSTs, where each reflects a custom affixation grammar per coarse lexical category (Section 4.2). The overall morphotactics model is obtained by composing those 15 FSTs.

5.1 Segmentation And Inflectional Groups

Following Hakkani-Tür et al. (2002) and Oflazer (2003), we segment a word into its root and inflectional groups (IG). IGs tokenize a word into subsegments based on the derivational boundaries that are in the word. As illustrated in Fig. 4 it is a complex segmental unit comprising of the derivational morpheme, lexical category of the derived form and inflections that might occur after that derivation.

In IG-based modeling last IG determines the final lexical category of the word and inflectional features of the last IG apply to the whole word in determining its grammatical function in sentence-level context. While building cascaded NLP architectures with lexemic syntactic processing units morphological features of the last IG are informative in PoS tagging and syntactic parsing to constraint data sparsity. We do not employ IG-based segmentation as a theoretical construct in our model, but rather include it as part of the morphological analysis representation. Together with IG boundaries we also represent segmentation of individual morphemes which is helpful in extracting morphemic grammars and assigning individual lexical categories to each morpheme.

5.2 Affix Inventory and Feature Tagset

Our affix inventory is composed of 51 inflectional and 72 derivational morphemes (excluding morphemes that are not realized in surface and by generalizing allophones over meta-phonemes). Inflectional morphemes are categorized over 8 feature categories (e.g. *Case* or *Possessive* on nominals, *Copula* or *TenseAspectMood* on verbals) and 42 feature values (e.g. *Case=Abl* or *TenseAspectMood=Aor*), whereas a single feature category is used to mark all derivations (*Derivation*) which can take 62 feature values (e.g. *Derivation=PastPart*). Compared to the models reported

(a) çaldığını ‘*that you stole (it)*’
(çal[VB]+[Polarity=Pos])([VN]-DHk[Derivation=PastNom]+[PersonNumber=A3sg] +SH[Possessive=P3sg]
+NH[Case=Acc])+[Proper=False]

(b) çaldığın ‘*(the thing) that you stole*’
(çal[VB]+[Polarity=Pos])([VJ]-DHk[Derivation=PastPart]+Hn[Possessive=P2sg]) +[Proper=False]

(c) koşarak ‘*(by) running*’
(koş[VB]+[Polarity=Pos])([CRB]-YArAk[Derivation=Ger])+[Proper=False]

Figure 6: PoS and derivational feature labeling for nominalizers, participles and converbials.

in the literature this is the most fine-grained morpheme segmentation model for Turkish.[4]

Çakıcı (2012) reports an affix inventory of 53 inflectional and 29 derivational morphemes, which is inherited from Oflazer et al. (1994) and used in extracting morpheme segmentations from MST. An investigation into the affix inventory of Şahin et al. (2013) shows that they do not represent some productive derivational processes (see Table 2). An example is -lA (Make), which creates denominal and deadjectival verbs in Turkish. According to Nakipoğlu and Üntak (2008) verbs derived by this suffix make up the largest portion of Turkish verb lexicon (excluding light verb constructions), accounting for about 21% of the verbs that are found in Turkish dictionaries. Çöltekin (2014) also does not represent -CAk (Coll), -CAnAk (Coll), -izm (Doct) -gil (Fam), -ist (Foll), -lA (Make), -lArcA (Of), -vari(Sim), -Hmtrak (Sim), -dA (Snd). Akın and Akın (2007) and Sak et al. (2009) does not segment infinitive markers from the root form. Sulubacak et al. (2016) consider verbal derivational morphemes -lAn (Acquire), -lAş (Become) and nominal derivational morphemes -CH (Agentive), -CHk and -CAğHz (Diminutive) on noun roots as a part of the root form, although they are semantically productive.

To represent the adequate phrasal scope of these affixes in morphemic syntactic processing and to recover clausal architecture in sentence-level disambiguation tasks in lexemic syntactic processing it is essential to explicitly segment and mark them. One example is Turkish subordination, which is handled through morphology. As illustrated in Fig. 3-b-d, noun, relative and adverbial clauses are created with an affix that attaches to the base verb to create a clause out of the sentence headed by the verb, which can then function as an argument or adjunct of the matrix verb.

Feature	Description	Morpheme	Example
Rcp	Reciprocal	-Hş	söyleş
Rfx	Reflexive	-Hn	süslen
Nonf	Nonfinite	-YHş	tüken-iş
Dim	Diminutive	-cAğHz	çocuk-cağız
Doct	Doctrine	-izm	komün-izm
Fam	Family	-gil, -lAr	annem-gil
Foll	Follower	-ist, -st	komün-ist
From	From	-lH	Ankara-lı
Lang	Language	-CA	Alman-ca
Ness	Ness	-lHk	insan-lık
Make	Make	-lA	işaret-le
Aff	Affinity	-CHl	et-çil
Of	Of	-lArcA	ton-larca
Sim	Similar	-Hmtrak, -vari	sarı-mtrak
Coll	Collective	-CAk, -CAnAk	toplu-canak
Ly	Adverbial	-CAsHnA	aptal-casına
Bcm	Become	-lAş	iyi-leş
Snd	Sound	-dA	fokur-da

Table 2: Derivational morphemes in our affix inventory distinct from Şahin et al. (2013).

We segment subordinating affixes that are described in Göksel and Kerslake (2004). They can be subcategorized into: (i) *Nominalizers* which create noun clauses (or verbal nouns), (ii) *Participles* which create adjectival clauses, (iii) *Converbials* which create adverbial clauses. A subset of these suffixal forms are ambiguous between two functions, they both create noun and adjectival clauses (e.g. -DHk affix in Fig. 3-b and Fig. 3-c). We explicitly mark differing functions of these in sentence-level context. Morphological feature tags for morphemes that create a noun clause end with -Nom (short for nominalizer, e.g. PastNom), and feature tags for those that create an adjectival clause end with -Part (short for Participle, e.g. PastPart). Words derived via attachment of subordinating affixes are also differentiated at the level of PoS. If they are derived by *Nominalizers* they receive the fine tag VN (verbal noun), words derived by *Participles* receive the tag VJ (verbal adjective) and those that are derived by *Converbials* are tagged as CRB (short for converbial). This brings in further syntactic expressivity to the morphological analyses as shown in Fig. 6.

[4] For an exhaustive list of morphemes segmented and tagged by our model, refer to `https://github.com/google-research/turkish-morphology/blob/master/analyzer/src/morphotactics/README.md`.

(a) ***İyi*** ile kötünün savaşı. ‘*The battle between the good and the bad.*’
(iyi[NN]+[PersonNumber=A3sg]+[Possessive=Pnon]+[Case=Bare]) +[Proper=False]

(b) 1000 liraya tablet baktım ama ***iyisini*** bulamadım.
‘*I have searched for a tablet to buy for 1000 liras but couldn’t find a good one.*’
(iyi[PRI]+[PersonNumber=A3sg]+SH[Possessive=P3sg]+NH[Case=Acc]) +[Proper=False]

(c) ***İyi*** bir insan. ‘*A good person.*’
(iyi[JJ])+[Proper=False]

(d) İlaç bana ***iyi*** geldi. ‘*The medicine made me feel well.*’
(iyi[RB])+[Proper=False]

(e) Bugün ***iyiyim***. ‘*I am well today.*’
(iyi[NOMP]+[PersonNumber=A3sg]+[Possessive=Pnon]+[Case=Bare] +[Copula=PresCop]
+YHm[PersonNumber=V1sg])+[Proper=False]

Figure 7: Morphological parses for root form *iyi* ‘*good*’ in 5 distinct sentence-level context.

5.3 Eliminating Zero-Derivation

The distinction between lexical categories is blurry in Turkish. Previous models employ a zero-derivation mechanism to capture this ambiguity, which is syntactic type shifting of a word through affixation of a so-called *empty morpheme* that does not realize in surface form. Instead, we cross-categorize lexical entries of root forms in the lexicon according to the syntactic functions they can take. This method ensures all derivational morphemes to have a corresponding realization in the surface. Representation-wise morphological parse ends up being significantly simplified and more tractable without empty morphemes while the base lexicon is kept compact and maintainable.

Fig. 7 presents disambiguated analyses for the word *iyi* ‘*good*’ in context. In its root form the word is 5-way ambiguous between categories NN, PRI, JJ, RB, and NOMP. As a preprocessing step prior to FST compilation such categorically ambiguous root forms are cross-categorized by adding new lexical items to the lexicon with a tag from the set of ambiguous lexical categories. We utilize a comprehensive set of cross-categorization rules that capture all ambiguous lexical category pairs.[5] This method enables us to strip lexical ambiguity handling from morphotactic model development while keeping morphotactic models for each lexical category generic. For example, word form *iyisi* (iyi+si, ‘good+SH[3Psg]’) will only be parsed as NN, NOMP, and PRI, where JJ and RB interpretations are pruned, even though the root form is cross-categorized for those tags. This is because the morphotactic model for JJ and RB would not allow root form *iyi* to be inflected for 3rd person possessive (P3sg).

5.4 Case Marking

Turkish is a nominative-accusative language where subjects are marked with nominative case (not realized in surface form) and direct objects with accusative (+YH and +NH). It is also shown to exhibit a grammatical phenomenon called *Head Incorporation*, which results in the verb forming a complex grammatical unit with its direct object (Kornfilt, 2003). In such cases direct object nominals do not have any case marking and they exhibit different behaviour from their cased counterparts in terms of syntactic and semantic properties.

Turkish is considered a free word order language where direct objects can be scrambled within the sentence from their canonical (preverbal) position (Bozşahin, 1998, 2000). However, as illustrated by Fig. 8-c caseless direct objects are less flexible to scramble and leave their preverbal positions.[6] Besides scrambling, caseless direct objects are also shown to be invisible to syntax in terms of binding and passivization (Aydemir, 2004; Öztürk, 2005, 2009). Furthermore, Aydemir (2004) shows that depending on whether the direct object has accusative case, the item that occurs before it can either be interpreted as an adjective or adverb. In Fig. 9-a, the noun *araba* has accusative marking, and modifier *iyi* is interpreted as an adjectival modifier of the noun. In Fig. 9-b, *araba* does not have any case and therefore invisible for syntactic modification, *iyi* is interpreted as an adverb and modifies the whole verb phrase. These

[5] For the complete set of cross-categorization rules that we use, refer to `https://github.com/google-research/turkish-morphology/blob/master/analyzer/src/morphotactics/README.md`.

[6] A detailed investigation into the extent of flexibility by which caseless objects can move from their preverbal positions is beyond the scope of this paper. For a thorough linguistic analysis, refer to Gračanin-Yüksek and İşsever (2011).

pieces of evidence are taken to indicate that caseless direct objects might not be forming syntactic arguments on their own.

(a)	Ahmet	dün	akşam	pasta	ye+di
	Ahmet	yesterday	evening	cake+Nom	eat+Past

(b)	Ahmet	pasta+yı	dün	akşam	ye+di
	Ahmet	cake+Acc	yesterday	evening	eat+Past

(c)	*Ahmet	pasta	dün	akşam	ye+di
	*Ahmet	cake	yesterday	evening	eat+Past

Figure 8: Scrambling, adopted from Kornfilt (2003).

(a)	Ahmet	iyi	arabayı	kullanır
	Ahmet	iyi	araba+YH	kullan+Hr

Ahmet[NNP]good[JJ]car[NN]+Accdrive[VB]+Aor
'*Ahmet drives the good car*'

(b)	Ahmet	iyi	araba	kullanır
	Ahmet	iyi	araba	kullan+Hr

Ahmet[NNP]good[RB]car[NN]+Baredrive[VB]+Aor
'*Ahmet drives well*'
(lit. '*Ahmet does good car-driving*')

Figure 9: Modification of caseless direct objects.

Previous Turkish morphology models mark such caseless objects with subjective case (nominative). They also extend application of subjective case to all other caseless nominals in the sentence, even to those that are caseless objects of postpositional phrases. We find this treatment syntactically problematic, because grammatical properties of subjects and caseless objects are completely different, so we label them distinctively. While a subject is marked with nominative case (Nom), caseless objects are marked to bear no case (Bare). These distinctive case features can be useful in downstream NLP tasks, especially in adequately disambiguating subjects from caseless objects in syntactic parsing.

5.5 Agreement

In Turkish a predicate agrees with its subject in Person and Number. As shown in Lewis (1967), Good and Alan (2000) and Göksel and Kerslake (2004) there are four suffixal paradigms for this agreement. The predicate can combine with affixes in one of these paradigms depending on its Tense-Aspect-Mood properties. Predicates having past tense (+YDH) or conditional (+YsA) are inflected with -k paradigm, those that are in imperative and optative mood are respectively inflected with imperative and optative paradigms, and all others are inflected with -z paradigm. Our model is sufficiently expressive of these paradigms based on agreement properties of predicates.

TenseAspectMood verbal inflectional feature that is marked on predicates clarifies which paradigm needs to be used in agreement morphology. Agreement itself is encoded in the *PersonNumber* feature of the morphological parse of verbals and nominals. Verbal agreement feature tags start with '*V*' prefix (e.g. V3sg), whereas for nominals prefix '*A*' is used (e.g. A3sg). Fig. 10 presents a scrambled raising construction, where embedded clause subject *seni* receives objective (accusative) case from the matrix verb *san*. Since the sentence is scrambled, word order is not a reliable indicator of which noun phrase is the subject of which clause. However, this information is easily recoverable from the morphological analyses using the agreement between *PersonNumber* features of the verbs and noun phrases.

5.6 Proper Nouns

We represent named entities as part of the morphological parse with the boolean feature *Proper*. All words that are part of a multi-word named entity are marked as *Proper*=*True*. This method allows us to label internal structure (PoS and morphological features) of multi-word named entities and spans of tokens that form them in sentence-level context (see Fig. 11). Trained over a representative corpus, a disambiguator based on such features of our model can output predictions whether a sequence of words form a named entity in context.

6 Testing and Evaluation

In order to test correctness of generated morphological analyses and identify possible gaps in the root form lexicon, we utilized a human-annotation based iterative development and testing scheme. 6 annotators, who are linguistically trained Turkish native speakers disambiguated morphological analyses that are output by our morphological analyzer by referring to sentence-level context. Annotation is done on a corpora of 2,200 sentences which are randomly extracted from Turkish Wikipedia pages. Annotators iteratively annotated batches of 200 sentences, reported illicit morphological analyses and word forms that cannot be parsed. Analyses for every word in the corpora is annotated by 2 annotators. The model and the root form lexicon is improved by taking account of syntactic constructions that are observed

Seni	ben	akıllısın	sandım
sen+YH	ben	akıllı+sHn	san+DH+m
you[PRP]+A2sg+Acc	I[PRP]+A1sg+Nom	smart[NOMP]+V2sg	consider[VB]+Past+V1sg
'*I considered you smart*'			

Figure 10: Person and number agreement in scrambled raising construction.

Yüzüklerin	Efendisini	izledim
Yüzük+lAr+NHn	Efendi+SH+NH	izle+DH+m
Ring[NN]+A3pl+Gen+**Proper=True**	Lord[NN]+P3sg+Acc+**Proper=True**	watch[VB]+Past+V1sg+**Proper=False**
'*I watched Lord of the Rings*'		

Figure 11: *Proper* feature labeling on named entities that span across multiple tokens.

in the corpora until no illicit analysis is reported and a satisfactory level of coverage is attained. Our improvements also aimed to refactor affixation grammars that are defined by the morphotactics model to limit overgeneration by disallowing affixation of certain derivational morphemes to a set of inflectional morphemes (e.g. -gil (Family) nominal derivation morpheme can only follow common and proper noun word stems or possessive inflections).

Table 3 shows coverage statistics of our model on a data set that is different than our development corpora. We define coverage as the fraction of word forms that our model can parse among the set of unique observed word forms. We calculate it over a merge of training and test set sentences of Turkish section of the CoNLL 2007 Shared Task of Dependency Parsing data set (Nivre et al., 2007), which contains 60,310 tokens and 18,443 unique word forms (after case-folding). On contrary to Çöltekin (2010) we do not remove tags, punctuation and numbers from the data set. The analyzer can parse 17,624 word forms, yielding 95.56% coverage, while generating 24.96 analyses and 2.06 IGs on average per word form. When we remove *Proper* morphological feature from the morphological parse, which generates duplicate analyses that only differ by this feature, the average number of analyses per word form is reduced to 12.82. Note that the coverage we report is not directly comparable with Şahin et al. (2013) since we do not employ any fallback mechanisms that depend on affix stripping. Such fallback methods potentially result in higher coverage with occasionally incorrect morphological parses.

Coverage statistics	
Tokens	60,130
Unique word forms	18,443
Accepted word forms	17,624
Unrecognized word forms	819
Coverage	95.56%
Average number of analyses	
With *Proper* feature	24.96
Without *Proper* feature	12.82
Average number of inflectional groups	
With *Proper* feature	2.06
Without *Proper* feature	2.05

Table 3: Statistics on analyzer coverage and average number of analyses and inflectional groups that it generates.

7 Conclusions and Future Work

In this paper we presented a syntactically expressive morphology model for Turkish, a human-annotated gold lexicon of root forms and a fine-grained affix inventory. While doing so, we also introduced a novel method to eliminate zero-derivations, a fine part-of-speech tagset and elaborate representations of inflectional/derivational features. We have shown that the implemented model has high coverage and does not overgenerate. In terms of lexemic syntactic processing, we would like to investigate implications of our representation in building morphological disambiguators and syntactic parsers. In parallel, we would also like to experiment with fully morphemic grammar induction, since our fine-grained morpheme segmentation scheme can be used in capturing adequate phrasal scope.

Acknowledgments

We would like to thank Martin Jansche for helping with Thrax and OpenFst integration; Haşim Sak for sharing their implementation; Ryan McDonald, Jan Botha and Bernd Bohnet for their comments; Savaş Çetin, Nihal Meriç Atilla, Eda Aydın Oktay, Faruk Büyüktekin, Hakan Keser, Hilal Yıldırım, İlmiye Karakimseli and Ozan Mahir Yıldırım for their meticulous work on data annotation.

References

Ahmet Afşin Akın and Mehmet Dündar Akın. 2007. Zemberek, an open source NLP framework for Turkic languages. *Structure*, 10:1–5.

Cyril Allauzen, Michael Riley, Johan Schalkwyk, Wojciech Skut, and Mehryar Mohri. 2007. OpenFst: A general and efficient weighted finite-state transducer library. In *International Conference on Implementation and Application of Automata*, pages 11–23. Springer.

Nart Bedin Atalay, Kemal Oflazer, and Bilge Say. 2003. The annotation process in the Turkish treebank. In *Proceedings of 4th International Workshop on Linguistically Interpreted Corpora (LINC-03) at EACL 2003*.

Yasemin Aydemir. 2004. Are Turkish preverbal bare nouns syntactic arguments? *Linguistic Inquiry*, 35(3):465–474.

Cem Bozşahin. 1998. Deriving the predicate-argument structure for a free word order language. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pages 167–173. Association for Computational Linguistics.

Cem Bozşahin. 2000. Gapping and word order in Turkish. In *Proceedings of the 10th International Conference on Turkish Linguistics*, pages 58–66.

Cem Bozşahin. 2002. The combinatory morphemic lexicon. *Computational Linguistics*, 28(2):145–186.

Ruket Çakıcı. 2005. Automatic induction of a CCG grammar for Turkish. In *Proceedings of the ACL student research workshop*, pages 73–78. Association for Computational Linguistics.

Ruket Çakıcı. 2012. Morpheme segmentation in the METU-Sabancı Turkish treebank. In *Proceedings of the Sixth Linguistic Annotation Workshop*, pages 144–148. Association for Computational Linguistics.

Çağrı Çöltekin. 2010. A freely available morphological analyzer for turkish. In *LREC*, volume 2, pages 19–28.

Çağrı Çöltekin. 2014. A set of open source tools for Turkish natural language processing. In *LREC*, pages 1079–1086.

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*.

Gülşen Eryiğit and Eşref Adalı. 2004. An affix stripping morphological analyzer for Turkish. In *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications*.

Aslı Göksel and Celia Kerslake. 2004. *Turkish: A comprehensive grammar*. Routledge.

Jeff Good and CL Alan. 2000. Affix-placement variation in Turkish. In *Proceedings of the 25th Annual Meeting of the Berkeley Linguistics Society: Special Session on Caucasian, Dravidian, and Turkic Linguistics*, pages 63–74.

Martina Gračanin-Yüksek and Selçuk İşsever. 2011. Movement of bare objects in Turkish. *Dilbilim Araştırmaları*, 22(1):33–49.

Dilek Z Hakkani-Tür, Kemal Oflazer, and Gökhan Tür. 2002. Statistical morphological disambiguation for agglutinative languages. *Computers and the Humanities*, 36(4):381–410.

Jaklin Kornfilt. 2003. Scrambling, subscrambling, and case in Turkish. *Word order and scrambling*, 125155.

Kimmo Koskenniemi. 1984. A general computational model for word-form recognition and production. In *Proceedings of the 10th International Conference on Computational Linguistics and 22nd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics.

Geoffrey L Lewis. 1967. *Turkish grammar*. Oxford University Press.

Mine Nakipoğlu and Aslı Üntak. 2008. A complete verb lexicon of Turkish based on morphemic analysis. *Turkic Languages*, 12:221–280.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan T McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *LREC*.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

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

Kemal Oflazer. 2003. Dependency parsing with an extended finite-state approach. *Computational Linguistics*, 29(4):515–544.

Kemal Oflazer, Elvan Göçmen, and Cem Bozşahin. 1994. An outline of Turkish morphology. *Report to NATO Science Division SfS III (TU-LANGUAGE), Brussels*.

Kemal Oflazer, Bilge Say, Dilek Zeynep Hakkani-Tür, and Gökhan Tür. 2003. Building a Turkish treebank. In *Treebanks*, pages 261–277. Springer.

Balkız Öztürk. 2005. *Case, referentiality and phrase structure*. John Benjamins.

Balkız Öztürk. 2009. Incorporating agents. *Lingua*, 119(2):334–358.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A Universal Part-of-Speech Tagset. In *LREC*.

Brian Roark, Richard Sproat, Cyril Allauzen, Michael Riley, Jeffrey Sorensen, and Terry Tai. 2012. The OpenGrm open-source finite-state grammar software libraries. In *Proceedings of the ACL 2012 System Demonstrations*, pages 61–66. Association for Computational Linguistics.

Haşim Sak, Tunga Güngör, and Murat Saraçlar. 2009. A stochastic finite-state morphological parser for Turkish. In *Proceedings of the ACL-IJCNLP 2009 Conference short papers*, pages 273–276. Association for Computational Linguistics.

Umut Sulubacak, Memduh Gökırmak, Francis Tyers, Çağrı Çöltekin, Joakim Nivre, and Gülşen Eryiğit. 2016. Universal dependencies for Turkish. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3444–3454.