Appendix: Turkish Morphological Features

In this section we provide a list of morphological features used in the encoding of about 9,000 possible IGs that can be produced by our morphological analysis. Although not all of these have been used in examples used in this chapter, we feel it is useful for conveying to the reader the wealth of the information Turkish lexical forms encode.

- **Major Parts of Speech**: `+Noun, +Adj, +Adv, +Conj, +Det, +Dup, +Interj, +Ques, +Verb, +Postp, +Num, +Pron, +Punc`. (`+Dup` category contains onomatopoeia words which only appear as duplications in a sentence.)
- **Minor Parts of Speech**: These typically follow a major POS to further subdivide that class, or to indicate the kind of derivation involved.
  - After `+Num`: `+Card, +Ord, +Percent, +Range, +Real, +Ratio, +Distrib, +Time`.
  - After `+Noun`: `+Inf, +PastPart, +FutPart, +Prop, +Zero`.
  - After `+Adj`: `+PastPart, +FutPart, +PresPart`.
  - After `+Pron`: `+DemonsP, +QuesP, +ReflexP, +PersP, +QuantP`.
- The following (mostly semantic) markers are used after derivations to indicate the kind of derivation involved:
  - After `+Adv` derived from verbs: `+AfterDoingSo, +SinceDoingSo, +As (he does it), +When, +ByDoingSo, +While, +AsIf, +WithoutHavingDoneSo`.
  - After `+Adv` derived from Adjectives: `+Ly` (equivalent to the English *+ly* derivation.)
  - After `+Adv` derived from temporal nouns: `+Since`
  - After `+Adj` derived from nouns: `+With, +Without +SuitableFor, +InBetween, +Rel`.
  - After `+Noun` derived from adjectives: `+Ness` (as in red vs. redness)
  - After `+Noun` derived from nouns: `+Agt` (someone involved in some way with the stem noun), `+Dim` (Diminutive),
  - After `+Verb` derived from nouns or adjectives: `+Become` (to become like the noun or adjective in the stem) `+Acquire` (to acquire the noun in the stem)
  - A `+Zero` appears after a zero morpheme derivation.

- Nominal forms (Nouns, Derived Nouns, Pronouns, Participles and Infinitives) get the following additional inflectional markers:

  1. **Number/Person Agreement**: `+A1sg, +A2sg, +A3sg, +A1pl, +A2pl, +A3pl`.
  2. **Possessive Agreement**: `+P1sg, +P2sg, +P3sg, +P1pl, +P2pl, +P3pl, +Pnon` (no overt agreement).
  3. **Case**:`+Nom, +Acc, +Dat, +Abl, +Loc, +Gen, +Ins`.

- Adjectives (lexical or derived) do not take any inflection, except `+Adj+PastPart` and `+Adj+FutPart` will have a `+Pxxx` (possessive agreement as above) to mark verbal agreement. Any other inflection to adjectives implies type-raising to nouns and the inflection goes onto the noun after a 0-morpheme derivation.

- Verbs have two sets of markers which are treated as derivations:

  1. **Voice**: `+Pass, +Caus, +Reflex +Recip`, (A verb form may have multiple causative markers).
  2. **Compounding/Modality**: `+Able` (able to verb), `+Repeat` (verb repeatedly), `+Hastily` (verb hastily), `+EverSince` (have been verb-ing ever since), `+Almost` (almost verb-ed but did not), `+Stay` (stayed frozen while verb-ing), `+Start` (start verb-ing immediately)

- Verbs also get the following inflectional markers:

  1. **Polarity**: `+Pos, +Neg`
  2. **Tense-Aspect-Mood**: A finite verb may have 1 or 2 of `+Past` (past tense), `+Narr` (narrative past tense), `+Fut` (future tense), `+Aor` (Aorist, may indicate habitual, present, future, you name it), `+Pres` (present tense, for predicative nominals or adjectives), `+Desr` (desire/wish), `+Cond` (conditional), `+Neces` (Necessitative, must), `+Opt` (optative, let me/him/her verb), `+Imp` (imperative), `+Prog1` (Present continuous, process), `+Prog2` (Present continuous, state).
  3. Verbs also have number person agreement markers (see nominal forms earlier) and an optional copula marker.

Notes

1. Literally, "(the thing existing) at the time we caused (something) to become strong". Obviously this is not a word that one would use everyday. Turkish words (excluding noninflecting frequent words such as conjunctions, clitics etc) found in typical text average about 10 letters in length.

2. Please refer to the comprehensive list of morphological features given in Appendix A for the semantics of some of the non-obvious symbols used here.

3. Though they may be separated by various clitics, in which case the collocation can not be recognized by simple local means.

4. This however does not mean that there no non-projective constructs in Turkish. There are a number of constructs, such as an adverbial modifying a verb, cutting in between a modifier and the head noun making up the subject NP. These, however, are very rare. Our representation does not have any restriction regarding projectivity and lets us represent the crossing links in such case.

5. Words in this context may also be a lexicalized or non-lexicalized collocations.

6. The input to the annotator is actually morphologically preprocessed with each token already having been analyzed in all its ambiguities. This same file could also be run through a morphological disambiguator module [7]. If this disambiguator makes any mistakes (and they do), our current tool does not let us correct an incorrectly disambiguated morphological analyses yet, so we have opted not to disambiguated for the time being.

References

[1] A. Abeillé, L. Clément, and A. Kinyon. Building a treebank for french. In A. Abeillé, editor, *Building and Exploiting Syntatically Annotated Corpora*, Text, Speech and Language Technology. Kluwer, 2001.

[2] A. Bémová, J. Hajič, B. H. J. Panenová, A. Böhmova, and E. Hajicova. The Prague Dependency Treebank. In A. Abeillé, editor, *Building and Exploiting Syntatically Annotated Corpora*, Text, Speech and Language Technology. Kluwer, 2001.

[3] T. Brants, W. Skut, and H. Uszkoreit. Syntactic annotation of a german newspaper corpus. In A. Abeillé, editor, *Building and Exploiting Syntatically Annotated Corpora*, Text, Speech and Language Technology. Kluwer, 2001.

[4] E. Erguvanlı. *The Function of Word order in Turkish*. PhD thesis, University of California, Los Angeles, 1979.

[5] J. Hajič. Building a syntactically annotated corpus: The Prague Dependency Treebank. In E. Hajicova, editor, *Issues in Valency and Meaning: Studies in Honour of Jarmila Panenova*. Karolinum – Charles University Press, Prague, April 1998.

[6] D. Z. Hakkani-Tür. *Statistical Language Modeling for Turkish*. PhD thesis, Bilkent University, Department of Computer Science, Ankara, Turkey, 2000.

[7] D. Z. Hakkani-Tür, K. Oflazer, and G. Tür. Statistical morphological disambiguation for agglutinative languages. In *Proceedings of COLING 2000*. ICCL, August 2000.

[8] J. Hankamer. Morphological parsing and the lexicon. In W. Marslen-Wilson, editor, *Lexical Representation and Process*. MIT Press, 1989.