## TURKISH LEXICAL SAMPLE TASK

Turkish is an interesting language that deserves being examined semantically. Effective parameters for WSD may vary for different languages and word types. Although, some parameters are common in many languages, some others may be language specific. Turkish is based upon suffixation, which differentiates it sharply from the majority of European languages, and many others. Like all Turkic languages, Turkish is agglutinative, that is, grammatical functions are indicated by adding various suffixes to stems. Turkish has a SOV (Subject-Object-Verb) sentence structure but other orders are possible under certain discourse situations. As a SOV language where objects precede the verb, Turkish has postpositions rather than prepositions, and relative clauses that precede the verb. As being one of the widely spoken languages over the world and being different from many other languages, Turkish is an appropriate language for semantic researches.

Turkish lexical sample task consists of three data. They are the dictionary, the training data, and the evaluation data.

### Dictionary

The dictionary is the one that is published by TDK (Turkish Language Foundation) and it is open to public via internet (http://tdk.org.tr/tdksozluk/sozara.htm). This dictionary lists the senses along with their definitions and example sentences that are provided for some senses. A typical entry from this dictionary for the word “şey (thing)” is given below:

**şey**
***isim Arapça şey¢***

**1 .** Madde, eşya, söz, olay, iş, durum vb.nin yerine kullanılan, belirsiz anlamda bir söz: *"Bana sen pek çok şey kazandırdın."*- R. H. Karay.

**2 .** Nesne, madde: *"Asıl zorluk belki öğrenilmesi lazım gelen şeylerin değil, unutulması gereken şeylerin çokluğundan gelir."*- A. Ş. Hisar

The entry in the dictionary has the following information: “**1 .** **(sense number)** Madde, eşya, söz, olay, iş, durum vb.nin yerine kullanılan, belirsiz anlamda bir söz **(definition)** *"Bana sen pek çok şey kazandırdın*."**(example sentence)**- R. H. Karay **(citation)**.” The dictionary is used only for sense tagging and enumeration of the senses for standardization. No specific information other than the sense numbers is taken from the dictionary, therefore there is no need for linguistic processing of the dictionary.

### Training and Evaluation Data

We will provide data for 35 words (10 nouns, 15 verbs and 10 other POS for the rest of POS including adjectives and adverbs). If a word has n senses, we tag at least 100 examples per word but the number of samples can be more depending on the n value. For a few words, however, fewer examples exist because of lack of data. In the final version, all the ambiguous words will have at least 100 examples. If for some words fewer examples exist in the corpus they can be either eliminated or some other examples can be added in the same format. At an average, each of the selected words have 10 senses, verbs, however, have more. Approximately 66% of the examples for each word will be delivered as training data, whereas 33% will be kept as evaluation data. Corpus samples will comprise 1-10 sentences including the target word depending on the Treebank files (i.e. corpus). Data will be given in ASCII files for each word under each POS. The samples for the words that can belong to more than one POS will be listed under the majority class. POS will be provided for each sample.

### Corpus

Lesser studied languages, such as Turkish suffer from the lack of wide coverage electronic resources or other language processing tools like ontologies, dictionaries, morphological analyzers, parsers etc. There are some projects for providing data for NLP applications in Turkish like METU Corpus Project (see here). It has two parts, the main corpus and the Treebank that consists of parsed, morphologically analyzed and disambiguated sentences selected from the main corpus, respectively. The sentences are given in XML format and provide many syntactic features that can be helpful for WSD. Treebank can be used for academic purposes by contract.

The texts in main corpus have been taken from different types of Turkish written texts published in 1990 and afterwards. It has about two million words. It includes 999 written texts taken from 201 books, 87 papers and news from 3 different Turkish daily newspapers. XML and TEI (Text Encoding Initiative) style annotation have been used. The distribution of the texts in the Treebank is similar to the main corpus. There are 6930 sentences in this Treebank. These sentences have been parsed, morphologically analyzed and disambiguated. In Turkish, a word can have many analyses, so having disambiguated texts is very important. Frequencies of the words have been found as it is necessary to select appropriate ambiguous words for WSD. There are 5356 different root words and 627 of these words have 15 or more occurrences, and the rest have less.

The sense tags are not included and have to be added manually. Sense tagging has been achieved for some words and hopefully, tags will be checked by some experts in order to obtain gold standard. Initial tagging process has been finished by a single tagger and controlled. Two other linguists in the team will tag and control the examples. That is, this step will be completed by three taggers. Problematic cases will be handled by a commission of three taggers that will act as the referee. The members of the commission will be different than the original three taggers and the decision will not be finalized until having 90% agreement in at most two months time.

The structure of the XML files (see Figure 3) contains tagging information in the word (morphological analysis) and sentence level (parse tree). In the word level, inflectional forms are provided (see Figure 1). And in the sentence level relations among words are given (see Figure 2). The S tag is for sentence and W tag is for the word. IX is used for index of the word in the sentence, LEM is left as blank and lemma is given in the MORPH tag as a part of it with the morphological analysis of the word. REL is for parsing information. It consists of three parts, two numbers and a relation. For example REL="[2, 1, (MODIFIER)]" means this word is modifying the first inflectional group of the second word in the sentence. The structure of the Treebank data was designed by METU. Initially lemmas were decided to be provided as a tag by itself, however, lemmas are left as blank. This does not mean that lemmas are not available in the Treebank; the lemmas are given as a part of "IG" tag. Programs are available for extracting this information for the time being. All participants can get these programs and thereby the lemmas easily and instantly.

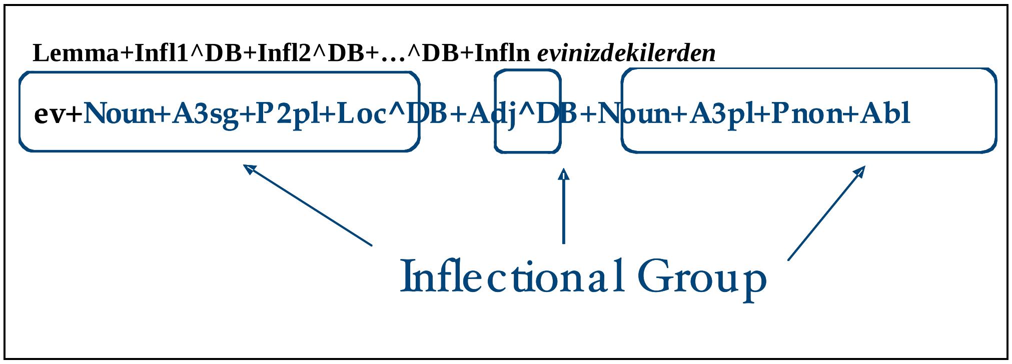

Lemma+Infl1^DB+Infl2^DB+…^DB+Infln *evinizdekilerden*
ev+Noun+A3sg+P2pl+Loc^DB+Adj^DB+Noun+A3pl+Pnon+Abl
Inflectional Group

**Figure 1: Inflectional groups for the Turkish word evinizdekilerden (from the ones in your house)**

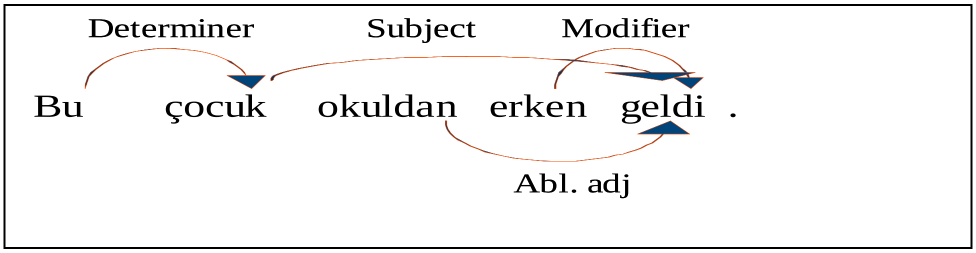
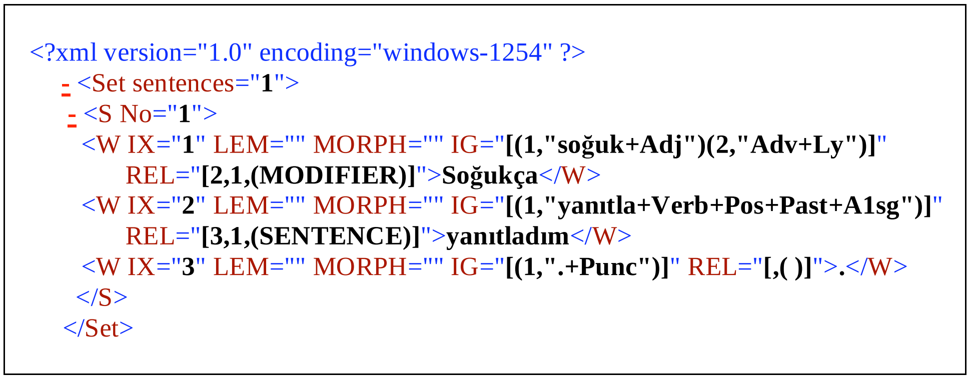


**Figure 2: Relational structure of Turkish sentence "Bu çocuk okuldan erken geldi" (This child came early from the school)**



```
<?xml version="1.0" encoding="windows-1254" ?>
  - <Set sentences="1">
   - <S No="1">
     <W IX="1" LEM="" MORPH="" IG="[(1,"soğuk+Adj")(2,"Adv+Ly")]"
          REL="[2,1,(MODIFIER)]">Soğukça</W>
     <W IX="2" LEM="" MORPH="" IG="[(1,"yanıtla+Verb+Pos+Past+A1sg")]"
          REL="[3,1,(SENTENCE)]">yanıtladım</W>
     <W IX="3" LEM="" MORPH="" IG="[(1,".+Punc")]" REL="[,( )]">.</W>
    </S>
  </Set>
```

**Figure 3: XML file structure of the Treebank**

We have extracted example sentences of the target word(s) and some features. Then ASCII files whose formats are complying with the arff file structure of WEKA system (see here) are obtained. The key files for each word are kept in ASCII format including information about Previous context (root, POS, inflected POS, case marker, possessor, relation), target word (root, POS, inflected POS, case marker, possessor, relation) and subsequent context (root, POS, inflected POS, case marker, possessor, relation). And feature files are also in ASCII format and are given below:

```
File id / Sentence# / Occurence# /Pre. Con. / Target/ Sub. Con.
0000221318.xml,1,0,biraz,adv,adv,null,fl,modıfıer, al,verb,verb,null,fl,sentence, punc,punc,punc,null,fl,null
```

The Treebank provides all necessary syntactical annotations. The sense tags are provided in the key files for each word. In the key files, sense annotations are given line by line. In each line file id, sentence# and occurrence# are given along with the fine-grained and coarse-grained sense of that specific word. One can use these key files and Treebank XML files to get any specific information about the word, context and the senses. These files for the training data will be open to public.

### Ontology

Small scale ontology for the target words and their context is still under construction and we are trying to provide this to the users. The Turkish WordNet developed at Sabancı University is somehow insufficient. Only the verbs have some levels of relations similar to English WordNet. This is not a suitable resource for fulfilling the requirements of Turkish lexical sample task. The ontology specific to this task is very close to the end. It will most probably be available before the end of the August. The ontology that is under construction for this task will cover the examples that are selected and will have two or three levels of relations (such as IS-A, HAS-A etc.) that are supposed to be effective in the disambiguation process.

### Format of the answers and Scoring

Format for answers (see here) and scoring (see Kilgarriff and Rosenzweig, section 8) is similar to SENSEVAL standard. A "guidelines to taggers" document, comprising detailed instructions of how instances were to be tagged and covering, e.g., multi-word units (including morphological and lexical variants on them), metaphors, missing meanings etc, will be made available with the lexical entries download if not before.