I can eat glass. It doesn't hurt.

Bill McNeill

May 8, 2004

The Facts of Farsi

The Farsi translation of the sentences "I can eat glass. It doesn't hurt" is as follows.

(1) *man* *mitavanam* *shishera* *bexoram*
I PROG-can-PRES-1SG glass-ACC SUBJ-eat-PRES-1SG
'I can eat glass.'

(2) *an* *dard* *nemikonad*
it harm NOT-PROG-make-PRES-3SG
'It doesn't hurt.'

The morphology of the verbs is as follows.

(3) *mi +* *tavan +* *am*
Progressive 'be able' Present 1st-person sigular
'I can'

(4) *be +* *xor +* *am*
Subjunctive 'eat' Present 1st-person sigular
'I may eat'

(5) *ne +* *mi +* *kon +* *ad*
Negative Progressive 'make' Present 3rd-person sigular
'I don't make'

The Farsi verb *tavanestand* means "to be able". It takes a verb phrase in the subjunctive as an object. The verb phrase must agree with the finite *tavanestand* in number and person. Although Farsi is in general an SOV language, *tavanestand* always comes before its verbal object.

Like most verbs in Farsi, "to hurt" is a compound formed by a auxiliary *kardan* that combines with a non-verbal element, in this case the noun *dard*. Dative pronomial objects of compound verbs appear as inflections on the non-verbal element. For example, "It doesn't hurt me" is *an dardam nemikonad*. I haven't implemented this in my grammar yet.

Farsi also allows subject pronouns to be omitted and usually only includes them for emphasis, so the above sentences are a little odd. A more natural way of saying this would be "*Mitavanam shishera bexoram. Dard nemikonad*." I haven't implemented a covert pronomial subject rule in my grammar yet.

Verbal negation is done with the prefix *ne-*. It must come before any other prefix.

For the moment, I am calling *mi-* a progressive prefix, but Farsi uses this "progressive" form for most present tense sentences. A tense without the *mi-* exists, but it is used in much more limited contexts. When we get around the doing the machine translation I'll probably change this to be just the present tense.

	Sentence	Description	Result
1	man mitavanam shishera bexoram	Grammatical	Pass
2	*man mitavanam shishera xoram	Controlled verb not subjunctive	Pass
3	*man mitavanam shishera mixoram	Controlled verb not subjunctive	Pass
4	*man shishera bexoram mitavanam	Controlled verb precedes tavanestan	Pass
5	*man mitavanam shishera bexorad	Controlled verb does match auxiliary person and number	Fail
6	*man mitavanam shishera	No controlled verb	Pass

Table 1: Regression tests for auxiliary verb

	Sentence	Description	Result
1	an dard nemikonad	"It doesn't hurt"	Pass
2	an dard mikonad	"It does hurt"	Pass
3	*an nemikonad	Missing compound object	Pass
4	*an mikonad dard	Compound object follows verb	Pass
5	*an dard minekonad	Incorrect prefix	Pass
6	*an dard benekonad	Incorrect prefix	Pass
7	*an dard mibekonad	Incorrect prefix	Pass
8	*an dard bemikonad	Incorrect prefix	Pass
9	*an dard nenekonad	Incorrect prefix	Pass
10	*an dard mimikonad	Incorrect prefix	Pass
11	*an dard bebekonad	Incorrect prefix	Pass

Table 2: Regression tests for compounds and verbal prefixes

"To be able" as a helping verb

I made *tavanestand* a raising auxiliary verb that requires a subjunctive verb on its COMPS list. Phrases headed by auxiliary verbs like *tavanestand* are generated by a head-initial head-comp-phrase rule, unlike the rest of my grammar's verb phrases, which are head-final. See the auxiliary-verb-lex rule for details.

Table (1) shows a suite of regression test sentences for the behavior of *tavanestand*. To pass, a grammatical sentence must be accepted by the parser and an ungrammatical sentence must be rejected. Test (5) fails because I haven't yet implemented the constraint that requires person and number agreement between *tavanestand* and its verbal object.

Negation and other prefixes

I implemented the compound verb head *kardan* with a rule that takes a non-accusative noun on the COMPS list. (I'm not sure if *kardan* also functions as a regular transitive verb in Farsi, so my grammar does not currently allow it to do so.) See the compound-verb-lex rule for details.

The negation, subjunctive, and progressive prefixes were all implemented as infl-ltol-rules. (They are ltol because the person and number suffix inflection is currently what I have triggering the conversion from words to lexemes.) By default all verbs are non-negated with completed aspect and indicative mood. The various verbal prefixes can change these values, and constraints in the inflectional rules ensure that prefixes combine in the correct order. See the progressive_lex_rule, subjunctive_lex_rule, and negation-lex-rule for details.

Table (2) shows a suite of regression test sentences for the behavior of compound verbs and verbal prefixes used in sentence (2) and elsewhere.

Semantics

Sentence (1) has the following MRS.

```
<h1,u2:SEMSORT,
{h3:pronoun_n_rel(x4:SEMSORT:FIRST:SG),
h5:pronoun_q_rel(x4, h7, h6),
h8:_can_v_rel(e9:SEMSORT:TENSE:ASPECT:MOOD, h10),
h11:_glass_n_rel(x12:SEMSORT:THIRD:NUMBER),
h13:def_q_rel(x12, h15, h14),
h10:_eat_v_rel(e17:SEMSORT:TENSE:ASPECT:MOOD, x16:FIRST:SG:SEMSORT, x12)},
{h6 qeq h3,
h14 qeq h11}>
```

There is one problem with this MRS, the relation eat takes the index x16 as its first argument instead of the first-person pronoun index x4. It appears that the raising verb is not semantically identifying its subject with the subject of the verb it controls. I'm not sure what's going wrong here.

Using the semantics of this sentence as a starting point, we overgenerate. We get back (1), but we also get back all possibile combinations of tense, aspect, and mood of *tavanestand* along with past and present versions of *xordan* "to eat", making twelve sentences total. For some reason information from the verb's TAM structure is not being entered into the semantic information, as is apparent from the MRS above. I'm not sure what's going wrong here.

Sentence (2) has the following MRS.

```
<h1,u2:SEMSORT,
{h3:pronoun_n_rel(x4:SEMSORT:THIRD:SG),
h5:pronoun_q_rel(x4, h7, h6),
h8:_harm_n_rel(x9:SEMSORT:THIRD:NUMBER),
h10:def_q_rel(x9, h12, h11),
h13:_make_v_rel(e14:SEMSORT:TENSE:ASPECT:MOOD, x4, x9),
h15:_neg_r_rel(u17:SEMSORT, h16)},
{h6 qeq h3,
h11 qeq h8,
h16 qeq h13}>
```

Because I'm using a compound verb, this looks slightly different from the MRS specified in the lab instructions. However, the correct relationships are being spelled out for the grammar as I have defined it. The predicate _make_v_rel is a two place predicate which takes the pronoun index x4 as its first argument and the _harm_n_rel index x9 as its second one. The predicate _neg_r_rel takes the handle h16 which is related via a qeq to h13, the handle for _make_v_rel.

As with (1), sentence (2) overgenerates by producing all possible combinations of tense, aspect, and mood on the main verb. It also generates versions of all the sentences with the animate *u* as the subject instead of the inanimate *an*. My grammar currently does not make a distinction between these two pronouns. Again there are twelve generated sentences total.