Building a Hybrid Machine Translation System for Translating from English into Persian

Fahime Mohammadpour (Corresponding author)
Faculty of Literature and Humanities, University of Sistan and Baluchestan
No. 19 Arash Alley, ArashStreet, Zahedan, Iran
Tel: 98-915-846-9037 E-mail: fahime_mohammadpour@yahoo.com

Abbas Ali Ahangar
Faculty of Literature and Humanities, University of Sistan and Baluchestan
English Language Department, Faculty of Literature and Humanities
University of Sistan and Baluchestan, Iran
Tel: 98-915-341-2856 E-mail:aaliahangar@yahoo.com

Nader Jahangiri
Ferdowsi University of Mashhad, Mashhad, Iran

Received: July 19, 2012 Accepted: August 13, 2012 Published: August 23, 2012
doi:10.5430/elr.v1n2p44 URL: http://dx.doi.org/10.5430/elr.v1n2p44

Abstract

The incredible speed of contemporary modern world has made using machine translation/ MT an essential need. This paper aims at describing a hybrid machine translation system conducted for translation of some simple declarative English sentences into Persian. It is called a hybrid machine translation system because it employs a combination of both rule-based approaches and corpus-based approaches. The results of the research suggest that the given machine translation system produces more natural sentences which are compatible with Persian word order.

Keywords: Machine translation system, Source language, Target language, Hybrid approach

1. Introduction

Translation is one of necessities of the today's fast world. To meet this requirement in a faster manner, the human being thought of machine translation. The short history of machine translation has witnessed employing three main approaches in designing MT systems: rule-based approaches, corpus-based approaches, and hybrid approaches.

Rule-based approaches and corpus-based approaches have been utilized in designing MT systems for translation from English into Persian more than hybrid approaches. Younesifar (1994) designed and built a machine translation system based on Augmented Transition Netwoks / ATN. The given system which used the sentence model of Wishon could produce a word for word translation of English sentences into Persian. Another machine translation system for translation from English into Persian is a rule-based machine translation based on HPSG. In this system, meaning is shown by Minimal Recursion Semantics / MRS semantic structure (Niknejad, 2008). Saedi (2008) introduced a hybrid machine translation for translation of simple English sentences into Persian. He (ibid) believed that disambiguation and transfer are the most important parts of this machine translation. Faroughi (2007) conducted a machine translation system based on lexical-functional grammar. The given system was designed for translating English sentences into Persian in general with a specific focus on translation of noun phrases. It seems that employing a hybrid approach would lead to generating better results.

This research aims to show that utilizing a hybrid approach, i.e. using LFG as its rule-based approach and using a statistical parser as its corpus-based approach, in building an MT system for translating of some English one-place, two-place, and three-place predicators into Persian will produce acceptable and natural sentences.

It is hoped that the findings of this research can be useful to translators, teachers, AI researchers and everyone else who needs an MT system.

Before giving a description of the hybrid MT system for translating some simple declarative sentences from English into Persian, the main approaches employed in designing MT systems will be reviewed and then lexical-functional grammar/LFG will be described. After that, the methodology will be illustrated. Finally, results will be presented.

1.1 Main Approaches Employed in Designing MT systems

Three main approaches have been employed in designing MT systems: rule-based approaches, corpus-based approaches, and hybrid approaches.

1.1.1 Rule-based Approaches:

Rule-based approaches includes: direct approach, interlingual approach, and transfer approach.

Generally, three approaches were used in MT systems before 1990s:

1) Direct Approach: In this approach, "systems were designed in all details especially for one pair of languages, i.e. in most cases, for Russian as SL and English as TL. The basic assumption was that the vocabulary and syntax of texts should be analyzed no more than necessary for the resolution of ambiguities, the correct identification of appropriate TL expressions and the specification of TL word order." (Hutchins, 1979: 31).

Aasi (2004: 34) believes that since, in this approach, translation process is done through replacement of the equivalent words, MT acts like a mechanical dictionary. MT systems which employ the direct approach are grouped as the first generation of MT systems.

The Georgetown University system, demonstrated in 1951, was typical of the 'direct' approach, because it "illustrates well the complexities and the ultimately insuperable problems of the 'direct' approach." (Hutchins, 1979: 31-32)

2) Interlingual Approach: This approach "assumes that it is possible to convert SL texts into semantic-syntactic representations which are common to more than one language. From such interlingual representations texts are generated into other languages. Translation is thus in two stages: from SL to the interlingua and from the interlingua to the TL." (Hutchins, 2003: 503) Hesabi (2006: 43) names this representation as "representation schema which is to some extent independent from SL and TL."

Hutchins (1979) assumes that Warren Weaver was the first one who mentioned the attractiveness of an interlingua approach to MT in his famous memorandum. "But it was not until the 1960's when theorical linguistics has turned to problems of language universals that MT researchers had any clear ideas of how interlinguals could be constructed." (Hutchins, 1979: 33)

Though Hutchins sees Interlingua approach positively, Wilks (2009:122) criticize this approach as: "an interlingual approach forces unneeded processing."

3) Transfer Approach: Experience with linguistically ambiguous MT systems which used interlingua approach led to the adoption of more modest 'transfer' approach.

Transfer approach includes three stages:

a) Transfer SL text to an SL-oriented abstract representation

b) Transfer the result representation to a TL-oriented abstract representation

c) Synthesis of TL text (Arnold, 2003).

These three approaches, i.e. direct approach, interlingual approach, and transfer approach, are called rule-based approaches.

1.1.2 Corpus-based Approaches:

Hesabi (2006: 46) cites " approaches which used abstract representation needed transferring all sentences to abstract representation, providing lexemes of all words used in translation of different languages, also process of analysis, transfer, and generation needed complex computation for each language separately. Moreover, processing speed of computers was slow in past times than now. Besides, there were some linguistic problems such as lexical and structural ambiguities, syntactic complexities, lexical and structural differences among languages, elliptical structures, and multi-word units such as idioms, which could not be translated correctly by computer. These all led to a situation in which computer experts decided to use approaches which did not need any abstract representation. Thus, rule-based MT systems changed to MT systems which employed corpus-based approaches."

There are two main types of corpus-based approaches: Example-based approaches, and Statistics-based approaches.

Hutchins (2005) describes them as following:

1) Example-based Approaches or Memory-based Approach:

"A system is EBMT system if it uses segments (word sequences (strings) and not individual words) of source language (SL) texts extracted from a text corpus (its example database) to build texts in a target language (TL) with the same meaning. The basic units for EBMT are thus sequences of words (phrases)." (Hutchins, 2005:63)

He identifies the basic processes of EBMT as: "the alignment of texts, the matching of input sentences against phrases (examples) in the corpus, the selection and extraction of equivalent TL phrases, and the adaptation and combining of TL phrases as acceptable output sentences."

2) Statistics-based Approach: In this approach, "SBM input was decomposed into individual SL words and TL words were extracted by frequency data (in the 'translation model') (Hutchins, 2005: 69).

MT systems which use one of these approaches belong to the third generation of MT systems. The distinctive feature of approaches employed in the third generation of MT systems with those employed in the first and second generation of MT systems is that corpus-based MT systems use no syntactic or semantic rules for analyzing text or equivalence selection. The comparison of the given definitions shows that final translation in EBMT is a sort of *translation by analogy* resulted from an analogy between SL and TL sentences. In contrast, SBMT systems uses statistically established word and phrase correspondences.

Some people such as Aasi (2004) are hopeful of corpus-based approaches and believe to advantages such as: no need to syntactic and semantic analysis, automation of all processes, and usability for other languages. But some others such as Somers (2001:143) are not that optimistic and say: "Time will tell if this is a major milestone in the history of machine translation, or just a minor diversion. Certainly there are plenty of researchers who prefer to continue exploring the conventional techniques, and already hybrid systems incorporating both approaches are also being reported."

1.1.3 Hybrid Approach:

This approach is a combination of both rule-based approaches and corpus-based approaches.

1.2 Lexical-Functional Grammar/LFG

In LFG, each sentence has three levels: Constituent Structure / C-structure, Functional Structure / F-structure, and Argument Structure / A-structure.

1.2.1 C-structure

Carnie (2007) believes that one major part of LFG is almost identical to the transformational grammar. "This is the idea that words of a sentence are organized into constituents, which are represented by a tree, and generated by rules" (Ibid: 438). Each tree represents C-structure of the given sentence. What makes the tree-diagrams in LFG different from the tree-diagrams in transformational grammar is the fact that in LFG there is no movement and no traces any more (Ibid).

In C-structure, relations such as dominance, precedence, and constituency are expressed through a series of phrase structure rules, which are represented in tree-diagram (Chatsiou, 2010).

1.2.2 F-stucture

Another structure in LFG is F-structure. F-structure "represents the relational structure of the sentence" (Van Valin, 2001).

Dabir Moghaddam (2004) believes that in F-structure, not only semantic information of each entry but also grammatical functions of the sentence constituents (i.e. subject, object, and verb) are represented. Moreover, grammatical functions such as subject and object are considered as nominal predicate. Also, for each noun, one 'specifier / SPEC' and one 'number / NUM' are recorded. Verbal predicate is consisted of a verb and its arguments (such as subject, object, subject complement, . . .). In F-structure, grammatical functions are called 'attribute' and their correspondence are called 'value'. So, such a representation is called 'Attribute Value Matrix / AVM'.

1.2.3 A-structure

Huang states that "C-structure and F-structure are language-dependent, that is they analyze the sentence based on linguistic information (segmental information) present in the sentence. But information related to the upper levels (i.e. supra segmental information), which are not present in the actual presentation of the sentence, should be considered to determine the meaning of the sentence. Thus, A-structure which presents the semantics information of the sentence is also paid attention to in LFG." (1993, cited in Faroughi, 2007).

Bresnan (1995, cited in Faroughi, 2007: 60) takes two parts for A-structure: Head or predicate and argument. In LFG, A-structure of a sentence shows the number of the participants in an event. Some of these arguments are obligatory and some are optional. This means that obligatory arguments cannot be deleted but optional arguments can be deleted.

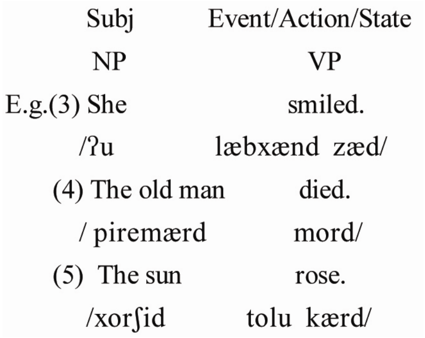
The A-structure of *put* is as below:

(2) *Put* 'place < (↑ SUBJ) (↑ OBJ) (↑ OBJ_{LOC}) >'

2. Method

Since the given MT system was designed for translating some of simple declarative English sentences -including one-place, two-place, and three-place predicators- into Persian, it is necessary to provide a contrastive analysis of them.

One-place Predicator: in this pattern, predicator denotes an action or state of the referent of the nominal argument (Yarmohammadi, 2002:59). In other words, this pattern includes predicators with a subject and no complement.

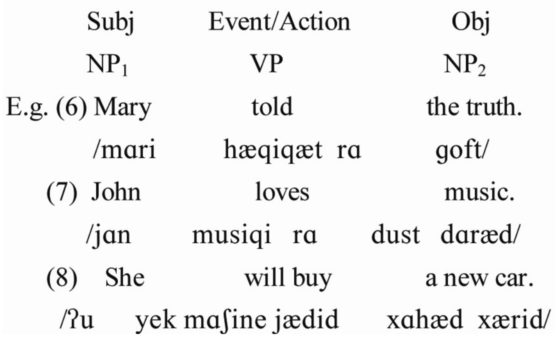


Subj Event/Action/State
NP VP
E.g.(3) She smiled.
/ʔu læbxænd zæd/
(4) The old man died.
/ piremærd mord/
(5) The sun rose.
/xorʃid tolu kærd/

As it is evident, there is a one-to-one correspondence between Persian and English for the above pattern.

Two-place Predicator: these are predicators which have a subject and one complement. The patterns of two-place predicators are as following:

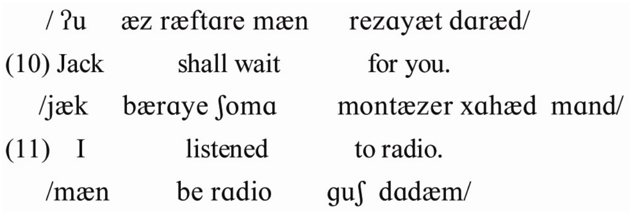
1. Predicator denotes an action or state which involves two objects - i.e., things or relationship between two objects (Yarmohammadi, 2002:78).



Subj Event/Action Obj
NP_1 VP NP_2
E.g. (6) Mary told the truth.
/mɑri hæqiqæt rɑ goft/
(7) John loves music.
/jɑn musiqi rɑ dust dɑræd/
(8) She will buy a new car.
/ʔu yek mɑʃine jædid xɑhæd xærid/

2. Predicator consists of a verb and a particle which come next to an object.

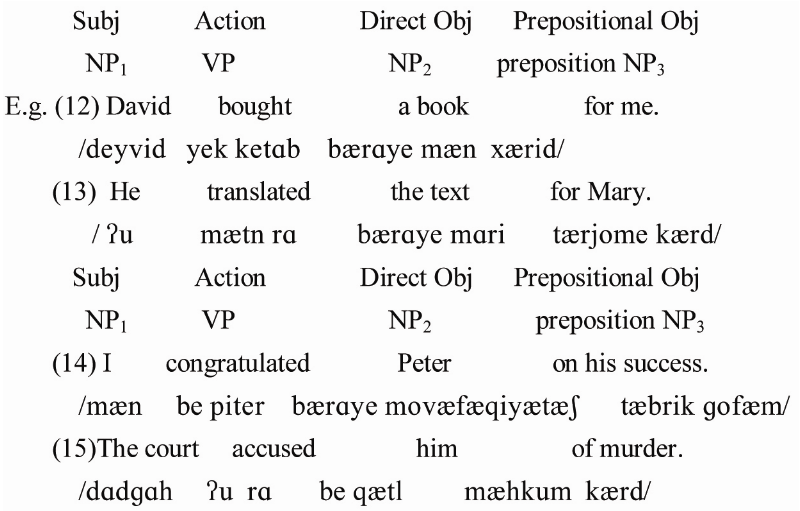
Subj Action Prepositional Obj
NP_1 VP NP_2
E.g. (9) He approve of my behavior.



/ ʔu æz ræftɑre mæn rezɑyæt dɑræd/
(10) Jack shall wait for you.
/jæk bærɑye ʃomɑ montæzer xɑhæd mɑnd/
(11) I listened to radio.
/mæn be rɑdio guʃ dɑdæm/

Three-place Predicator: these are predicators which need three complements. They are as follows:

1. Predicator involves a subject, a direct object, and a prepositional object.



Subj Action Direct Obj Prepositional Obj

NP_1 VP NP_2 preposition NP_3

E.g. (12) David bought a book for me.

/deyvid yek ketɑb bærɑye mæn xærid/

(13) He translated the text for Mary.

/ ʔu mætn rɑ bærɑye mɑri tærjome kærd/

Subj Action Direct Obj Prepositional Obj

NP_1 VP NP_2 preposition NP_3

(14) I congratulated Peter on his success.

/mæn be piter bærɑye movæfæqiyætæʃ tæbrik gofæm/

(15)The court accused him of murder.

/dɑdgɑh ʔu rɑ be qætl mæhkum kærd/

2.1 Data Analysis:

First of all, some simple English sentences -including one-place predicators, two-place predicators, and three-place predicators- were randomly selected from English grammar books –such as Advanced Grammar in Use and Modern English- to be analyzed in MT. The data analysis in this MT system comprises three stages: analysis of source language (i.e. English) sentences, selecting suitable equivalence, and generating target language (i.e. Persian) sentences.

2.1.1. Analysis of Source Language Sentences:

The selected English sentences were parsed by a statistical parser named "Statistical Parsing of English Sentences". This parser was able to parse input sentences and draw a tree-diagram for each of them. It could also determine the part of speech of words of a sentence and syntactic category of its noun phrases. This is what happens in C-structure of LFG.

What makes drawing the tree-diagram of input sentences an essential step in data analysis is the fact that determination of part of speech of words in tree-diagram will be helpful in cases of lexical ambiguity. A word such as *spring* may be a noun or a verb. As a noun it means "the season between winter and summer" and as a verb it means "to jump". In cases like this, the suitable equivalence could be selected if the parser had determined the part of speech of the given word properly.

To simulate what happens in F-structure of LFG, some algorithms were written to determine the grammatical function of noun phrases present in tree-diagram. These algorithms were based on the following simple sentences:

- Subject is NP daughter of S,
- (Direct) object is NP daughter of VP,
- Prepositional object is NP daughter of PP (Carnie, 2002:79).

Therefore, the grammatical functions of noun phrases in sentence (16) were determined as follows: *we* as subject, *the spaghetti* as direct object, and *the fork* as prepositional object.

2.1.2 Selecting the suitable equivalence:

After determination of part of speech of words and grammatical function of noun phrases present in a sentence, the next stage is to select the suitable equivalence for each of words and phrases of a sentence. To accomplish this purpose, a dictionary consists of basic words of English which made the translation of simple sentences possible, was prepared.

2.1.3 Generating target language sentences:

At this stage, the MT system put the selected equivalence of each word or phrase next to each other based on Persian word order, that is:

subject, (direct object), (prepositional object), verb

It is worth mentioning that object marker in Persian language is *ra* which appears after object word in a sentence. So, adding object marker *ra* was considered while writing algorithm.

A part of the algorithm of the given MT system is what follows:

```
hparse = parse;
            string participaltense="";
            if (FindTokenParticipal(hparse, ref participaltense) != "not participal")check participal sentence
            {
                TheSentenceIsParticipal = true;
                result = ParticipalTranslate(hparse, FindTokenParticipal(hparse, ref participaltense), participaltense);
            }
            for (k = 0; k < parse.ChildCount; k++)
            {
                parse2 = parse.GetChildren()[k];
                if (parse2.Type == "S" || parse2.Type == "SBAR") //this is for infinitive sentence
                {
                    Toparse=parse2.GetChildren()[0];
                    hparse = parse2;
                    if (FindTokenInfinitive(parse2) == true)
                    {
                        TheSentenceHasInf = true;
                        FindKindOfSubjectForInfinitive(parse2);
                        result = InfinitiveTranslate(hparse);
                    }
                    else
                        result = Translation(hparse);
                    finalresult = finalresult + " " + result;
                }
                else if (parse2.Type == "VP")
 {
```

3. Results and Conclusion

English one-place predicators, two-place predicators, and three-place predicators were translated into Persian as one-place predicators, two-place predicators, and three-place predicators correspondingly. While generating target language stage, taking into consideration the Persian word order made the generated sentences more natural for people whose mother tongue is Persian.

The following sentences were given to the system for translation:

(17) She coughed.

ترجمه : او سرفه کرد.

(18) We may go.

ترجمه : ما شاید برویم.

(19) The child fell.

ترجمه: کودك افتاد.

(20) I laughed.

ترجمه: من خندیدم.

(21) Mary saw her.

ترجمه: ماری او را دید.

(22) The sun melted the ice.

ترجمه: خورشید یخ را ذوب کرد.

(23) Peter kicked the ball.

ترجمه: پیتر توپ را با پا زد.

(24) The manager will explain the problem.

ترجمه: مدیر مسئله را توضیح خواهد داد.

(25) The earthquake destroyed the city.

ترجمه: زلزله شهر را ویران کرد.

(26) The children may enjoy the film.

ترجمه: کودکان از فیلم شاید لذت ببرند.

(27) He was speaking to me.

ترجمه: او با من مشغول حرف زدن بود.

(28) Tom gave a flower to her.

ترجمه: تام گل را به او داد.

(29) They named her Jane.

ترجمه: آنها او را جین نامیدند.

(30) My mother bought me a shirt.

ترجمه: مادرم برای من یک بلوز خرید.

(31) His parents will write a letter to his teacher.

ترجمه: والدینش یک نامه برای معلم او خواهند نوشت.

(32) He repeated the sentence for the child.

ترجمه: او جمله را برای کودک تکرار کرد.

It is worth mentioning that the research was done on very small samples and the results of this study are confined to the English, as source language, and Persian, as target language. So, the same research on two other languages may produce different results. To generalize results, more studies must be done.

This paper aimed to show that a hybrid machine translation system built with employing a combination of both rule-based approach –i.e. LFG- and corpus-based approach –i.e. a statistical parser- for translation of some simple declarative English sentences ,including one-place predicators, two-place predicators, and three-place predicators, is able to generate acceptable sentences into Persian. To accomplish this purpose, at first stage, a tree-diagram was drawn for each of the sentences. At the second stage, the suitable equivalence was selected for each of the determined words and phrases. Finally, target language sentences were produces based on Persian word order.

The findings of this research suggest that this hybrid MT system is able to determine the part of speech of words of a sentences and their syntactic category correctly. Moreover, using a bilingual dictionary for selecting the suitable equivalence is of prime importance for generating more comprehensible and natural translation of English sentences into Persian.

The question that may be raised is whether this system is able to produce a natural translation of other types of English sentences into Persian.

References

Aasi, M. (2004). Machine Translation: From Dream to Reality. In F. Farahzad (Eds.), *Proceeding of two translation studies conferences* (pp. 31-48). Tehran: Yalda Ghalam.

Arnold, D. (2003). Why translation is difficult for computers. In H. Somers (Eds.), *Computer and Translation: a translator's guide.* USA: John Benjamin's North America.

Bresnan, J. (1995, October). *Lexicality and argument structure*. The Paris syntax and semantics conference, Paris. Retrieved from http:// www.lfg. Stanford.edu/lfg/archive/archive

Carnie, A. (2007). *Syntax: A Generative Introduction* (2nd ed.). USA: Blackwell Publishing.

Chatsiou, A. (2010). *An LFG Approach to Modern Greek Relative Clauses.* PhD dissertation. Department of Language and Linguistics, University of Essex.

Dabir Moghadam, M. (2004). *Theorical Linguistics: Emergence and Development of Generative Grammar* (2^{nd} ed.). Tehran: Samt.

Faroughi, J. (2007). *A Context-based Model for Machine Translation of Some Simple Sentences from English to Persian based on Lexical-Functional Grammar*. Mashhad: Ferdowsi University of Mashhad dissertation.

Hesabi, A. (2006). Abstract Representation in Machine Translation. *Translation Studies, 13*, 41-48.

Huang, C. R. (1993). *Mandarin Chinese and the lexical mapping theory: a study of the interaction of Morphology and argument changing*. The Bulletin of the institute of History and Philology.

Hutchins, J. (1979). Linguistic models in machine translation. *UEA Papers in Linguistics.* Retrieved from http://www. hutchinsweb.me.uk

Hutchins, J. (2003). Machine Translation: General Overview, by Mitkov, Ruslan. *The Oxford Handbook of Computational Linguistics*, 501-511. Oxford: Oxford University Press.

Hutchins, J. (2005). *Towards a definition of example-based machine translation*. Retrieved from http://www. hutchinsweb.me.uk

Niknejad, S. A. (2008). *Designing and Building a Machine Translation from Persian into English.* Tehran: Sharif University of Technology M.A. Thesis.

Saedi, Ch. (2008). *Machine Translation of Persian Sentences into English: A Hybrid Approach.* Tehran: Science and Research Branch Islamic Azad University M.A. Thesis.

Somers, H. (2001). Machine translation: Application. In M. Baker (Ed.), *Routledge encyclopedia of translation studies*. London and New York: Routledge.

Van Valin, R. (2001). *An Introduction to Syntax*. USA: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139164320

Wilks, Y. (2009). *Machine Translation: Its Scope and Limits*. UK: Springer.

Yarmohammadi, L. (2002). *A Contrastive Analysis of Persian and English (Grammar, Vocabulary and Phonology)*. Tehran: Payam Noor University.

Younesifar, F. (1994). Designing and Conducting a Machine Translation system based on Syntactic ways. Tehran:

Sharif University of Technology M.A. Thesis.

Notes

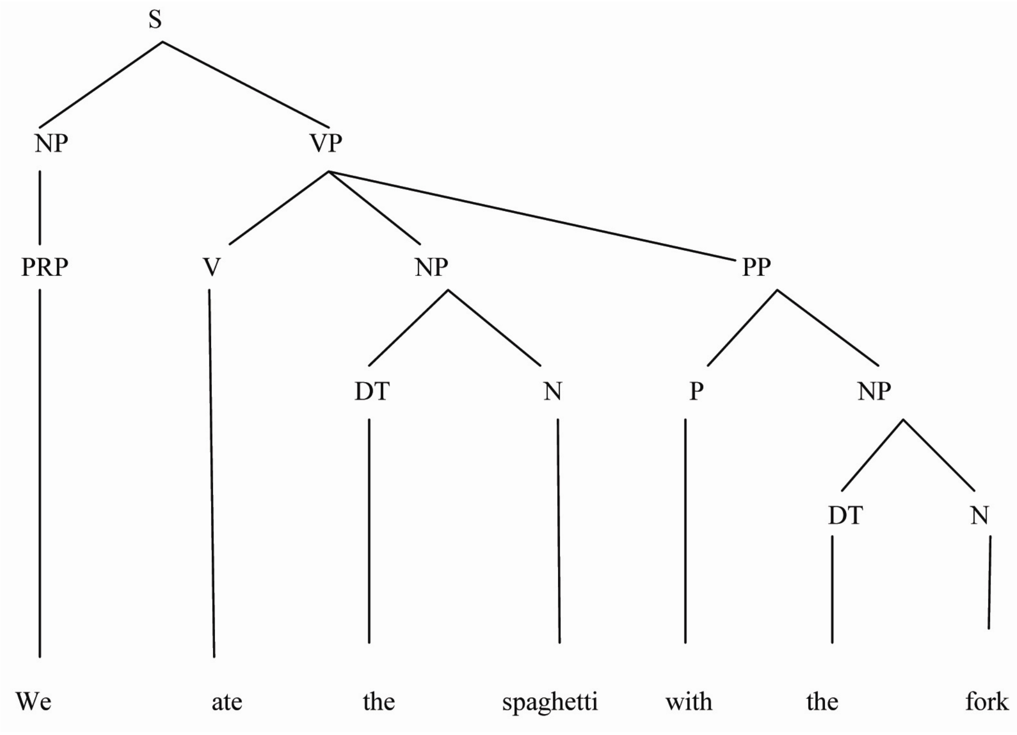
The F-structure of the sentence (1) is as follows:

(1) The man hid the truth.

$$\left[\begin{array}{ll} \text{PRED} & \text{‘hid < SUBJ, OBJ>’} \\ \text{TENSE} & \text{past} \\ \text{SUBJ} & \left[\begin{array}{ll} \text{DEF} & + \\ \text{NUM} & \text{sng} \\ \text{PRED} & \text{‘the man’} \end{array}\right] \\ \text{OBJ} & [\,\text{PRED} \quad \text{‘the truth’}] \end{array}\right]$$

The tree-diagram of the sentence (2) is as following:

(2) We ate the spaghetti with the fork.


S
NP
VP
PRP
V
NP
PP
DT
N
P
NP
DT
N
We
ate
the
spaghetti
with
the
fork