Formal Model of Adjective in The Kazakh Language

A.Mukanova
asel_ms@bk.ru

B.Yergesh
b.yergesh@gmail.com

A. Sharipbay
sharalt@mail.ru

G.Bekmanova
gulmira-r@yandex.kz

L.N Gumilyov Eurasian National University

Abstract

This paper explains how semantic hypergraphs are used to construct ontological models of morphological rules in the Kazakh language. The nodes within these graphs represent semantic features (morphological concepts) and the edges within represent the relationships between these features. Word forms within the hypergraph structure are described in trees which are converted into linear parenthesis notation; the trees and the linear parenthesis notations correspond to each other. Linear parenthesis notations are the formal models of morphological rules and the software implementation of the linear parenthesis notation allows for the automation of the synthesis of the various morphological word form analyses of the Kazakh language.

Introduction

Agglutinative languages (lat. Agglutinatio — combine, stick) are languages that have a system in which the dominant type of inflection is the agglutination ("sticking") of different formants; these can be either a prefix or a suffix and have only one meaning [1].

The Kazakh language is part of the Turkic group of languages; this language group can be classified as an agglutinative language. Words in the Kazakh language contain many word inflections; inflections are formed by adding suffixes and endings to words. Suffixes and endings are attached in a strict sequence and words in the Kazakh language vary in number, case, and person. A possessive form in Kazakh exists as it does in the English language [2-3].

Currently, ontology is a powerful and widely used tool which is used to model the relationship between objects of different subject fields. It is acceptable to classify ontology based on the degree of dependence on the task or application area, the model of ontological knowledge representation and expressiveness as well as other parameters [4]. Applied ontologies describe concepts which depend on both the task and the subject field of ontology.

Applied ontology is based on the general principles of ontology building, using semantic hypergraphs as a model for the representation of knowledge. This formalism will determine ontology O as triplet (V, R, K) where V is a set of concepts of the subject field (hypergraph nodes), R is a set of relationships between these concepts (hypergraph and edges), and K is a set of the names of concepts and relationships in the given subject field.

The semantic hypergraph language is a formal means of the representation of knowledge in which it is possible to implement classifying, functional, situational, and structural networks and scenarios, depending on the relationship types. This language is an extension of semantic networks where N-ary relations are represented naturally; these relations not only

Gönderme ve kabul tarihi: 10.09.2014-25.10.2014

allow for the specification of the objects' attributes but also permit a representation of their structural, "holistic" descriptions [5].
There are some papers on the use of semantic hypergraph [6-7]. Zhen L, Jiang Z. [7] describes the semantic hypergraph model as a 'hyper-graph based semantic network' (Hy-SN), which can represent more complex semantic relationships and which have a more efficient data structure for storing knowledge in repositories.
In [8-9] the hypergraph H (V, E) is defined by the pair (V, E), where V is the set of vertices $V = \{v_i\}$, $i \in I = \{1, 2, \ldots, n\}$, and E is set of edges $E = \{e_j\}$, $j \in J = \{1, 2, \ldots, m\}$; each edge is a subset of V. Vertex v and edge e is described as an incident if v $\in$ e. For v $\in$ V by d(v) denotes the number of edges incident to a vertex v; d(v) is called the degree of a vertex v. Degree of edge e, the number of vertices incident to this edge, is denoted by r(e).
Use of the ontology model for the representation of morphological rules allows for the translation of the morphological model on an almost one to one basis within the object-oriented data model. Where classes are the part of speech of the Kazakh language and the objects refer to their semantic categories, for example, qualitative type, relative type and degrees of comparison of adjectives.
Use of the ontology model for the representation of morphological rules part of speech allows describing complete morphological model with their relationships. Use semantic hyper graph for the representation of morphological rules part of speech and structure (frame) for the representation the concept. This representation allows translating to the object-oriented data model, where semantic hypergraph vertices are classes.
The purpose of this research is the automated generation of word forms and new words in the Kazakh language as well as the morphological analysis of the Kazakh language.
The research problem consists of the difficulties of formalizing of any natural language.
The authors believe that the problem of formalization of the Kazakh language is handled well through the proposed model below.
In this paper we describe an adjective in Material and Method section. The formal model of a noun is described in [10].

2 Material and Method

The semantic features of the initial forms of adjective (Adj) are qualitative (qual), relative (rel) and comparison (comparison); the sign determines the trajectory of the inflection of the adjective. Adjective in the Kazakh language conjugate (pers_end) and variesy for case (cases), as well as numbers (number) and have a possessive form (poss_end), comparison degree.
We used the ontology editor Protege [11] to build an ontology. It is a free and open source ontology editor and framework for building knowledge bases and is being developed at Stanford University in collaboration with the University of Manchester. Figure 1 shows the ontological model of adjective with its semantic features.

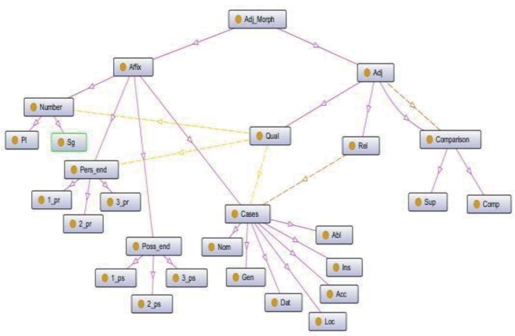

Figure1. Ontological model of adjective

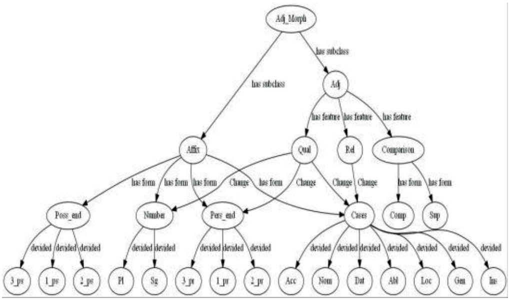

Figure2 . Visualization of adjective as a graph

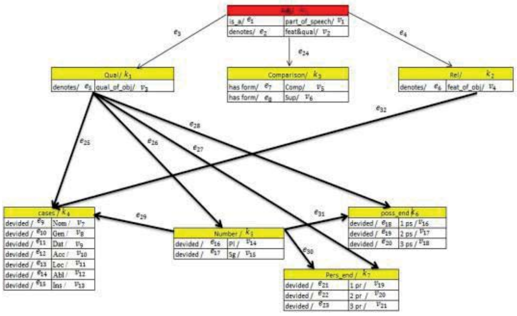

Figure3 . Graphical representation of ontology using semantic hypergraph

Table 1 describes the concepts and relationships used in the ontology

ID	Notation	Description
k_0	Adj	Adjective
v_1	Part_of_ speech	Part_of_ speech
v_2	Feat&qual	Quality and feature
k_1	Qual	Qualitative
v_3	Qual_of_obj	Quality of object
k_2	Rel	Relative
v_4	Feat_of_obj	Feature of object
k_3	Comparison	Comparison
v_5	Comp	Comparative
v_6	Sup	Superlative
k_4	Cases	Cases
v_7	Nom	Nominative case
v_8	Gen	Genitive case
v_9	Dat	Direction-dative case
v_{10}	Acc	Accusative case
v_{11}	Loc	Locative case
v_{12}	Abl	Ablative case
v_{13}	Ins	Instrumental case
k_5	Number	Number
v_{14}	Pl	Plural
v_{15}	Sg	Singular
k_6	Poss_end	Possessive endings
v_{16}	1 ps	1 personal
v_{17}	2 ps	2 personal
v_{18}	3 ps	3 personal
k_7	Pers_end	Personal endings
v_{19}	1 pr	1 personal
v_{20}	2 pr	2 personal
v_{21}	3 pr	3 personal
e_1	is_a	is a
e_2	denotes	Denotes
e_3, e_4, e_{24}	has_feature	has feature
e_5, e_6	has	Has
e_7, e_8	has form	has form
$e_9 - e_{23}$	devided	Devided
$e_{27} - e_{28}$ e_{32}	change	Change
$e_{27}, e_{30},$ e_{31}	add	Add

Hyper-arcs will be called as semantic arcs for separating semantic hypergraphs from other types of graphs; it will also be assumed that the set of vertices of the semantic hypergraph includes set of classes $K = \{k_a\}$, where $a \in A = \{0, 1, 2, 3..., n\}$ each of which will consist of set of instances of the class [12]. Thus, vertex-class can be represented by triple:

$k_a = \{V_a, E_a, S_a\}$,

where V_a - set of class properties, E_a - set of semantic arcs incident to class, S_a - set of instance of class.

The adjective vertex-classes:

$$k_0 = \{\{v_1, v_2\}, \{e_1, e_2\}, S_0\}$$
$$k_1 = \{\{v_3\}, \{e_5\}, S_1\}$$
$$k_2 = \{\{v_4\}, \{e_6\}, S_{2'}\}$$
$$k_3 = \{\{v_5, v_6\}, \{e_7, e_8\}, S_{3'}\}$$
$$k_4 = \{\{v_7, v_8, v_9, v_{10}, v_{11}, v_{12}, v_{13}\},$$
$$\{e_9, e_{10}, e_{11}, e_{12}, e_{13}, e_{14}, e_{15}\}, S_4\}$$
$$k_5 = \{\{v_{14}, v_{15}\}, \{e_{16}, e_{17}\}, S_{5'}\}$$
$$k_6 = \{\{v_{16}, v_{17}, v_{18}\}, \{e_{18}, e_{19}, e_{20}\}, S_6\}$$
$$k_7 = \{\{v_{19}, v_{20}, v_{21}\}, \{e_{21}, e_{22}, e_{23}\}, S_7\}$$

We can represent the adjective's morphological model with the semantic hypergraph model:

Hypergraph H(V,E), where

$$V = K = \{k_a\}, E = \{e_a\}$$
$$V = \{k_0, k_1, k_2, k_3, k_4, k_5, k_6, k_7\}$$
$$E = \{e_3 = \{k_0, k_1\}, e_4 = \{k_0, k_2\},$$
$$e_{24} = \{k_0, k_3\}, e_{25} = \{k_1, k_4\},$$
$$e_{26} = \{k_1, k_5\}, e_{27} = \{k_1, k_6\},$$
$$e_{28} = \{k_1, k_7\},$$
$$e_{29} = \{k_5, k_4\}, e_{30} = \{k_5, k_7\},$$
$$e_{31} = \{k_5, k_6\}, e_{32} = \{k_2, k_4\}\}$$

3 Results and Discussion

We have the base of initial forms containing 40,000 words with semantic features. Here 5,000 words are adjective. From the above described semantic hypergraph we can obtain formal rules using the parenthesis notation. The number of formal rules for adjective are 2,000. Through the use of these formal rules 65,000 word forms of the adjective are generated; it is also possible to generate adjective from other parts of speech.

As an example the inflection of the adjective word "akyldy" (in english "clever") includes all word forms of this adjective and their morphological information, which in abbreviated notation contains information on which case of the adjective, and which person is an action and whether it belongs to one or another person. An example shows the inflection of the adjective "akyldy" in cases.

Example. Inflection of the adjective "akyldy"

S=akyldy

$$k_0 = \{\{v_1, v_2\}, \{e_1, e_2\}, \text{akyldy}\}$$
$$e_4 = \{\text{akyldy}, \text{Rel}\}$$
$$e_{32} = \{\text{Rel}, \text{Cases}\}$$
$$k_4 = \{\{v_7, v_8, v_9, v_{10}, v_{11}, v_{12}, v_{13}\},$$
$$\{e_9, e_{10}, e_{11}, e_{12}, e_{13}, e_{14}, e_{15}\}, S_4\}$$

{akyldy(ақылды), akyldynyn(ақылдының), akyldyga(ақылдыға), akyldyny(ақылдыны), akyldyda(ақылдыда), akyldydan(ақылдыдан), akyldymen(ақылдымен)}

On the basis of these rules the morphological analyzer for the Kazakh language was created. It can be used to create spell checking technology of the Kazakh language and can be a cornerstone for translators, semantic search engines, speech technologies, etc.

Many methods of formalizing the morphological rules of a natural language do not allow the description of the semantic properties of words. This paper elaborates on the possibility of using semantic hypergraphs as a tool in order to formalize the morphological rules of any natural language based on the semantic features of words. Although this paper uses the Kazakh language to illustrate this concept the semantic hypergraph can be applied to any natural language.

Earlier results were obtained using a semantic neural network. 2.8 million word forms were generated from 40,000 initial word forms; these results were approved in [13]. The application of the semantic hypergraph allowed an increase of the number of word forms to 400,000 units. This was achieved by a complete description of the semantic features of words, which utilized the expressive power of the semantic hypergraph.

In the future we plan to apply this proposed method towards other Turkic languages.

4 Conclusion

The construction of ontological models of the morphological rules of Kazakh language allowed for the creation of formal rules of inflection and word formation for each part of speech. Software implementation of these rules made it possible to automatically generate more than 3.2 million word forms (dictionary entries) from 40,000 initial word forms with marked semantic features

5 References

[1] **EİFRİNG H., THEİL R.** (online), Linguistics for Students of Asian and African Languages. Available at: http://www.uio.no/studier/emner/hf/ikos/EXFAC03-AAS/h05/larestoff/linguistics/ (accessed 19/08/2014)

[2] ***KAZAKH GRAMMAR***. Phonetics, word formation, morphology, syntax, Astana, 2002. In Kazakh.

[3] ***BATAYEVA Z.,*** Colloquial Kazakh, Routledge, 2012.

[4] ***GRUBER, T.R.,*** Toward Principles for the Design Of Ontologies Used for Knowledge Sharing, International Journal Human-Computer Studies, 1995, Vol. 43, Issues 5-6, p. 907-928.

[5] ***KHAKHALİN, G.,*** Applied Ontology in the language of hypergraphs, Proceedings of IInd All–Russian Conference "Knowledge - Ontology - Theory" (KONT-09), 2009, p. 223-231. In Russian.

[6] ***RUİTİNG LİAN, BEN GOERTZEL, SHUJİNG KE, JADE O'NEİLL, KEYVAN SADEGHİ, SİMON SHİU, DİNGJİE WANG, OLİVER WATKİNS, GİNO YU,*** Syntax-Semantic Mapping for General Intelligence: Language Comprehension as Hypergraph Homomorphism, Language Generation as Constraint Satisfaction, Artificial General Intelligence. Lecture Notes in Computer Science , 2012,Volume 7716, p. 158-167.

[7] ***ZHEN, L., JİANG, Z.***, Hy-SN: Hyper-graph based semantic network, Knowledge-Based Systems, 2010, Vol 23, Issue 8, p. 809-816.

[8] ***BRETTO A.,*** Hypergraph Theory, Springer International Publishing Switzerland, 2013.

[9] ***BERGE, C.C.***, Graphs and Hypergraphs, Elsevier Science Ltd., 1985.

[10] ***BANU YERGESH, ASSEL MUKANOVA, ALTYNBEK SHARİPBAY, GULMİRA BEKMANOVA, AND BİBİGUL RAZAKHOVA,*** Semantic Hyper-graph Based Representation of Nouns in the Kazakh Language. Computación y Sistemas Vol. 18, No. 3, 2014 pp. 627–635, ISSN 1405-5546, DOI: 10.13053/CyS-18-3-2041

[11] ***PROTÉGÉ.*** Available at: http://protege.stanford.edu (accessed 19/08/2014)

[12] ***POTCHİNSKII I.***, Formal representation of semantic hypergraphs and their operations, 2012. Available at: http://rgu-penza.ru/mni/content/files/2012_Pochinskii.pdf

[13] ***SHARİPBAEV A.A. , BEKMANOVA G.T., BURİBAYEVA A.K., YERGESH B.Z., MUKANOVA A.S., KALİYEV A.K.***, Semantic neural network model of morphological rules of the agglutinative languages, 6th International Conference on Soft Computing and Intelligent Systems, and 13th International Symposium on Advanced Intelligence Systems, SCIS/ISIS, 2012, p. 1094-1099.