THE SEMANTICAL, ONTOLOGICAL MODELS AND FORMALIZATION RULES KAZAKH COMPOUND WORDS

L.Zhetkenbay, A.A.Sharipbay, G.T.Bekmanova, M.Khabylashimuly, U.Kamanur

L.N. Gumilyov Eurasian National University, Astana, Kazakhstan

jetlen_7@mail.ru, sharalt@mail.ru,gulmira-r@yandex.ru, mu_la@mail.ru, unzila.88@mail.ru

ABSTRACT

This thesis describes the rules of formulation of compound words in Kazakh language and their ontological, semantic models.

Key words. *Compound words, ontology, semantic hyper-graphs.*

1 Introduction

In modern linguistics the compound words are considered to be the most complicated problem because it has lots of unclear and vague aspects. It is difficult to define a language which doesn't have the compound words, but according to the written facts we can reveal that the compound words are widely used especially in German, English, Japan, Hindi, Russian languages. The role of compound words are different in any language. According to it's role in the lexicology of Japan language, it can be even called as the language of compound words. And such language family as Turkic languages which include also Kazakh language have rich vocabulary of the compound words. Scientists are saying that the compound words are very antic phenomenon in the Turkish languages and they can be found in the written heritage of the Orkhon memorials which give some visual examples from the names of the people and landmarks. The compound words like other derivative words are the final results of word construction. Among Kazakh scholars this issue was taken into consideration by K.Zhubanov for the first time. Even in the early 1930 years, he made a conclusion that compound words have a single meaning in spite of being composed of two or more parts and give a name to a particular object, play the role of one part of a sentence. But some of them change their external appearance as a result of the sound modification of one part of a compound word and they can be put with only one stress on a particular syllable. The compound words are composed of two or more words. Thus it's main characteristic is a compound construction and the main reason for differentiating them from the simple words[1].

Nowadays the formalism is the most significant issue in the process of representing natural language. So the compound words are the main complicated issue in the linguistics. That's way we need to determine the formalism as a way to solve this issue and semantic hyper-graphs could be considered as a one of them.

Graph-based models are widely used to represent natural languages [2].

Ontology is a powerful and widely used tool to model relationships between objects belonging to various subject fields. It is possible to classify ontologies based on the degree of dependence on the task or application area, the model of ontological knowledge representation and expressiveness, as well as other criteria [3]. Applied ontologies describe concepts that depend on both the task and the subject domain of the ontology.

An applied ontology is based on general principles of ontology building, and semantic hyper-graphs are used as a model for knowledge representation. This formalism determines ontology O as triple (V, R, K), where V is a set of concepts of a given subject field (hyper-graph vertex), R is a set of relationships between these concepts (hyper-graph edges), and K is a set of names of concepts and relationships in the domain. The semantic hyper-graph language is a formal means of knowledge representation, 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 allow us to specify the attributes of objects, but also permit representing their structural, "holistic" descriptions [4, 5].

Hyper-graphs generalize standard graphs by defining edges between multiple vertices instead of only two vertices [5]: a hyper-graph H = (V; E) on a finite set $V = \{v_i\}_{i \in I}$, $I = \{1, 2, ..., n\}$, of vertices is a family $E = \{ej\}_{j \in J}$, $J = \{1, 2, ..., m\}$, of subsets of V called hyper-edges; I and J are finite sets of indexes. Hyper-graphs are plural constructions. But sometimes it's better to use singular ones. That's way we need to transform the information contained in the plural constructed hyper-graphs to the singular constructed hyper-graphs. In order to design graphs as singular construction we can use such symbols like lines and brackets, so it means to design hyper-graphs as lines. There the tops of the hyper-graphs are designed as elliptical brackets. There are very close connection between the hyper-graphs and their line-brackets design, because they could be designed in such a way only after thorough investigation. The examples for graphs and their corresponding line-brackets design are shown at the 1 Figure.

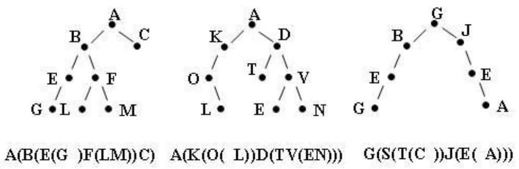


Figure-1. A model for line-brackets ways of writing.

2 Models and formalization rules Kazakh compound words

The Kazakh language has 5 types of compound words. They are: converged words, united words, joint words, abbreviated words, word combinations. Below there is shown ontological models of compound words and the rules of formalizing them by using line-brackets models of writing. Such rules are used to create the compound words of Kazakh language. Figure 2 and 3 show the ontological model of creating the noun category of the Kazakh language according to the structure. The compound words are differentiated according to their structure as following: converged words (buegin, zhaezdigueni, byjyl), united words (baspasoez, eltangba, elbasy, Temirqazyq), joint words (uelken-kishi, ajaq-tabaq), abriviated words (ENU, USA, UNN), word combinations (tang namaz, qara torghaj, qara koek, zhuez bes).

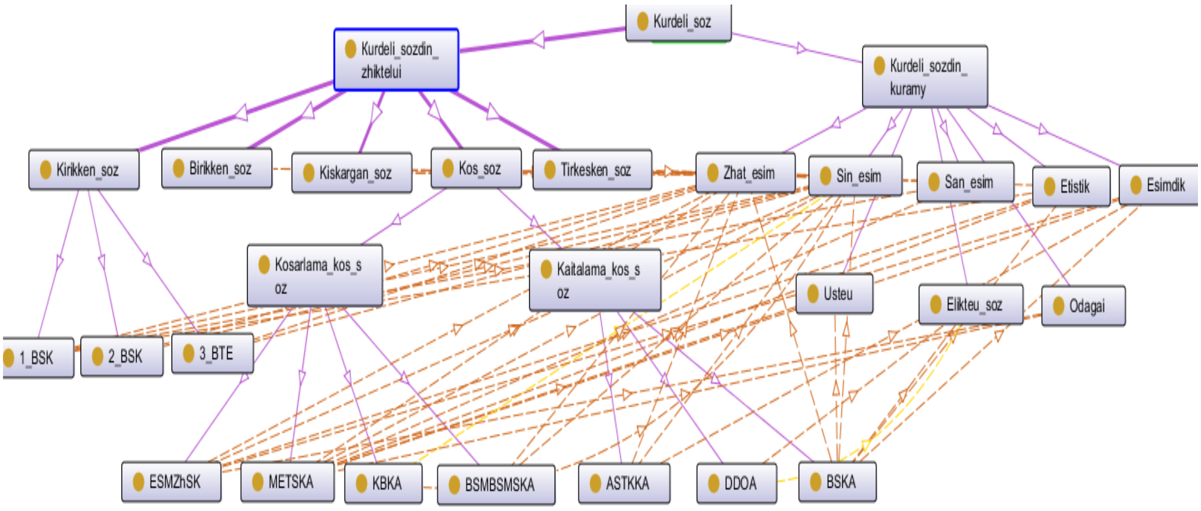


Figure-2. The ontological model of Kazakh compound words by their types

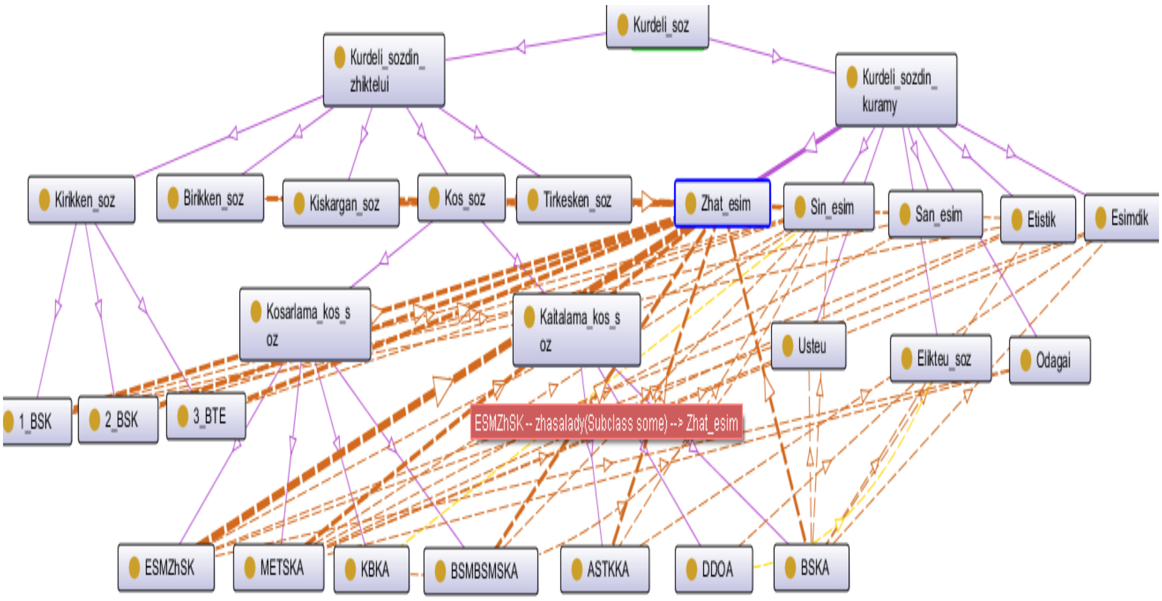


Figure-3. The ontological model of words made of nouns by their structure

We used the ontology editor Protégé (http://protege.stanford.edu) to build the ontology. It is a free open source ontology editor and a framework for building knowledge bases. It was developed at Stanford University in collaboration with the University of Manchester. Figure 2, 3 shows the ontological model of the Kazakh Compound words.

We made this ontology according to the compound words in Kazakh language and their types. The highest classification of them we called 'Kurdely soz' and there are two lower calssifications as 'Kurdely sozdin zhiktelui' and 'Kurdely soz kuramy'. So the classification 'Kurdely sozdin zhiktelui' include: converged words, united words, joint words, abbreviated words, word combinations. And as their lower classification we decided to include the rules of creating compound words.

As for the lower classification of 'Kurdeli soz kuramy' we took: the nouns, adjectives, numerals, verbs, pronouns, adverbs, exclamations. In order to connect them we invented a function 'Zasalady'and it has two possibilities. Like, the both parts of compound words could be of noun words or adjectives, e.c.c. For example, both parts of the joint word 'ata-ana' are nouns, but the one part of the word dunyekumar 'dunye' is the noun and the second part 'kumar' is an adjective. So we have made a model for each compound word.

Now we will make some rules concerning Kazakh compound words and to formalize them by the way of brackets.

The condition: first we need to systemize the letters and let's indicate them as following.

Latin letters (http://e-zerde.kz/latin/index2.php)

AOUYE!01
AeOeUeIEJa IjJu!02
MNNg!03 // silent vowel letter
MNNg!04 // loud vowel letter
RWJ!05
RWJ!06
L!07
L!08
BGGhD!09

BGGhD!10
ZhZ!11
ZhZ!12
P!13
P!14
K!15
K!16
Q!17
STSh!18
STSh!19

The rule 1. By adding the consonant letter ‘p’ (п) to an adjective beginning with a vowel letter we can create a compound word. Taking into consideration the above mentioned condition:

01X! ((01p)-X) - ap-alasa (the very low)
01X! ((01p)-X) - ap-aryq (the very thin)
01X! ((01p)-X) - yp-ystyq (the very hot)
02X! ((02p)-X) - aep-aedemi (the pretty)
02X! ((02p)-X) - uep-uelken (the very big)

The rule 2. By adding the consonant letter ‘p’ (п) after the first two letters to an adjective beginning with a consonant letter we can create a compound word.

Taking into consideration the above-mentioned condition:

0301X! ((0301p)-X) - mop-momyn (the very modest)
1002X! ((1002p)-X) - bip-bijik (the very high)
1101X! ((1101p)-X) - zhap-zhasyl (the very green)
1202X! ((1202p)-X) - zhip-zhingishкe (the very thin)
1202X! ((1202p)-X) - zhep-zhengil (the very easy)
1602X! ((1602p)-X) - koep-кoene (the very old)
1701X! ((1701p)-X) - qap-qara (the very black)
1801X! ((1801p)-X) - sap-sary (the very yellow)
1902X! ((1902p)-X) - taep- taeti (the very sweet)

The rule 3. By repeating a single word can also be made a compound word (repeated joint word).

X! ((X)-X) - bir-bir (one)
X! ((X)-X) - tez- tez (fast)
X! ((X)-X) - taw-taw

The rule 4. By adding to the imperative verbs can be made a compound word (by adding -a, -e, -j (-а, -е, -й) particles to the verbs).

((X06))! (((X06)e)-X06(e))- Кoere-кoere (having repeatedly seen)
((X06))! (((X06)e)-X06(e))-zhuere- zhuere (gradually)
((X05))! (((X05)a)-X05(a))- bara-bara (over time)
((X06))! (((X06)e)-X06(e))-zhuegire-zhuegire (run)
((X08))! (((X08)e)-X08(e)) - кuele-кuele
((X02))! (((X02)e)-X02(e))-soeilej-soeilej

БГҒД!10
ЖЗ!11
ЖЗ!12
П!13
П!14
К!15
К!16
Қ!17
СТШ!18
СТШ!19

The rule 5. By repeating a single word and by adding to it’s first part the particles: -pa, -pe, -ta, -te, -ma, -me, -ba, -be, -da, -de (-па, -пе, -та, -те, -ма, -ме, -ба, -бе, -да, -де), can be made a compound word. Respectively, if a word ends in rigid consonants should be added -па, -пе, if a word ends in loud consonants -ba, -be, and if a word ends in a sonorous consonant -ma, -me, -da, -de, particles can be added.

((X19))! (((X19)pe)-X) - betpe-bet (face to face)
((X06))! (((X06)de)-X) - birde-bir (any)
((X11))! (((X11)ba)-X)–awyzba- awyzba (confidentially)
((X12))! (((X12)be)-X) - zhuezbe-zhuez (face to face)
((X07))! (((X07)ma)-X) - qolma-qol (at once)
((X08))! (((X08)me)-X) – daelme-dael (literally)

The rule 6. By combining two words and the first component ends in a vowel, the last component also has vowel phoneme, as a result the last vowel sound of the first word drops out for making a compound word.

$$\begin{cases} Y_1 + Y_2 = Y \\ X_1 + X_2 = X \end{cases}$$

If we show the system by the adding method: $Y_1 + Y_2 + X_1 + X_2 = Y + X$

Here, Y_2, X_1 - vowel sounds

Alma+aty=Almaty
ala+aiaq=alaiaq
Qara+aspan=Qaraspan
qara+ala=qarala
Qandy+aghash=Qandaghash

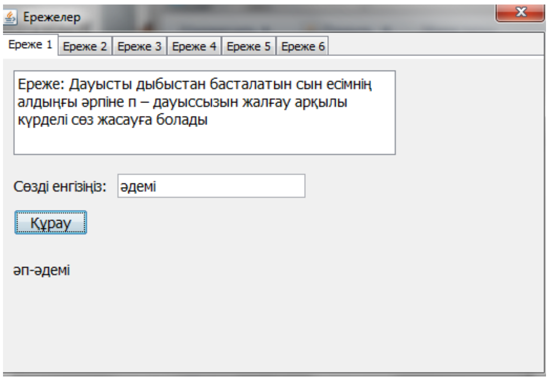
3 Results

In the basis of semantic network the ontological model of Kazakh compound words were built and rules of compound words were formed. In short, in the basis of semantic network the semantic model of Kazakh compound words were focused. By the rules, Kazakh compound words formed, automated and the program were made the algorithm of program were fulfilled by Net beans program in Java language.

There was made and collected an information base more than 1600 semantically featured compound words in kazakh.

Compound words in Kazakh were represented through graphs and was made the ontological model of compound words in Kazakh language.

As the result of the research, there was made a new base of compound words in Kazakh and the rules of formation for the compound words and its algorithm. On the the bases of the algorithm the creation of compound words was automated. For proceeding with the work we should enter in the box "enter a word" a word according to the "Rule 1" (Figure 4).



Ереже: Дауысты дыбыстан басталатын сын есімнің алдыңғы әрпіне п – дауыссызын жалғау арқылы күрделі сөз жасауға болады

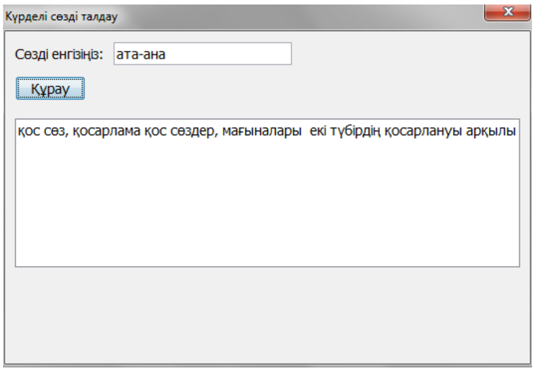
Сөзді енгізіңіз: әдемі

әп-әдемі

Figure-4. The rule 1

As we see above, after having entered the word "әдемі", we make the command to "construct". And according to the "Rule 1" there will be made a new compound word. So by applying above mentioned 6-rules we can make any compound words automatically.

The program for making semantic models, constructing and analyzing Kazakh compound words was indicated in the Figure 5. In the box "Enter a word" we should write a compound word, and by choosing the command "construct", the program will select the exact compound word and analyze it.



Сөзді енгізіңіз: ата-ана

қос сөз, қосарлама қос сөздер, мағыналары екі түбірдің қосарлануы арқылы

Figure-5. A window for analyzing a compound word

4 Conclusion

In this thesis we made ontological model of Kazakh language, the rules of formalizing of Kazakh compound words. As a working language of the thesis we used semantic hyper-graphs. As a result we revealed the set of rules which automatically made the compound words. There was made a data base describing semantic characteristics of more then 1600 kazakh compound words. These rules and formalisms could be applied for another different systems, like machine translation, various searching programs.

5 References

[1] Kazakh grammar. (2002). Phonetics, word formation, morphology, syntax (in Kazakh). Astana.

[2] Martins, R. (2012). Knowledge Vertices in XUNL. Polibits, Vol. 45, 61–66.

[3] 15 Gruber, T.R. (1995). Toward Principles for the Design of Ontologies Used for Knowledge Sharing. International Journal Human

[4] Khakhalin, G. (2009). Applied Ontology in the language of hypergraphs (in Russian). Proceedings of 2nd All–Russian Conference "Knowledge—Ontology—Theory" (KONT-09). Novosibirsk, 223–231.

[5] Liu, H., LePendu, P., Jin, R., & Dou, D. (2011). A Hypergraph-based Method for Discovering Semantically Associated Itemsets. In ICDM 2011, 398–406.