

Metalanguage and Knowledgebase for Kazakh Morphology

Gaziza Yelibayeva(✉), Assel Mukanova, Altynbek Sharipbay, Altanbek Zulkhazhav, Banu Yergesh, and Gulmira Bekmanova

L.N. Gumilyov Eurasian National University, Astana, Kazakhstan
gaziza_y@mail.ru, asiserikovna@gmail.com, sharalt@mail.ru, altinbekpin@gmail.com, b.yergesh@gmail.com, gulmira-r@yandex.kz

Abstract. Currently, the volume of various information resources in the Turkic languages is increasing. Processing of such resources requires thesauri and corpora created using a single metalanguage (tagging language) and the knowledge base of subject areas. This article proposes a meta-language of the morphological concepts of the Turkic languages on the example of the Kazakh language, which was used to create the knowledge base of the morphology of the Kazakh language in the Protégé environment. The results of the work were used to develop software applications for semantic search and knowledge extraction, as well as an assessment of knowledge applications on the morphology of the Kazakh language in the e-learning system.

Keywords: Kazakh language · Knowledge representation · Knowledge extraction · Metalanguage · Morphological rules · Ontology · Protégé

1 Introduction

To process natural languages, it is necessary to create a tagging system. The existing metalanguages mainly contain the concepts of the Roman-Germanic and the Slavic language groups. These metalanguages are not suitable for the description of the Turkic languages, which have many concepts different from the mentioned language groups. Therefore, the creation of the UniTurk meta-language for tagging up texts of the Turkic languages, including the Kazakh language, is an urgent task. Such a metalanguage will allow to unify tagging, facilitate their understanding and use common software, and conduct various studies on linguistic-statistical comparative analysis among the Turkic languages [1–3].

The metalanguage is needed to create a common resource with which all the developers of the Turkic electronic corpora could work. Such a resource could serve as a reference system for both developers and users of the Turkic electronic corpora. The most appropriate components of such a resource that meet the required conditions are ontological models of the grammar of the Turkic languages.

It is known that ontologies are used as data sources for many natural language processing applications, they allow to process complex and diverse information in a

© Springer Nature Switzerland AG 2019
S. Misra et al. (Eds.): ICCSA 2019, LNCS 11619, pp. 693–706, 2019.
https://doi.org/10.1007/978-3-030-24289-3_51

more efficient way. When managing knowledge of subject areas, ontological models are applied at the structuring stage and are considered as special knowledge bases. An ontology can act as a knowledge base framework, that is, can create a framework used to describe key concepts related to a specific subject area, and can formally be represented as the following triple [4, 5]:

$$O = <X, R, F> \tag{1}$$

where X is a finite set of concepts (terms) of the domain, which is represented by the ontology O; R is a finite set of relationships between the concepts (terms) of a given domain; F is a finite set of interpretation functions defined on concepts or relations of ontology O. The role of the interpretation function can be played by a verbal explanation of a term (annotation), a formula for calculating the meaning of a term, an algorithmic description, and also a definition in the form of a logical formula.

Ontological modeling of the morphological concepts of the Kazakh language was done in the Protégé environment, which allows simplifying the process of creating, downloading, changing and transforming the knowledge base, and to provide it for general use for joint viewing and editing [6, 7].

2 Related Works

The metalanguage is designed for tagging up the texts of natural languages and by using its means, it is possible to express all that is expressible by means of the object language and to designate all signs, expressions of the object language, which have names. In the metalanguage, one can express the properties of the object language expression and the relations between them. Here one can formulate definitions, notation, rules of formation and transformation for the objective language expressions [8].

There are well-known tagging systems, such as Penn Treebank [9] and CLAWS tagging system [10], which is used for tagging the British Corpus [9], and the tagging system of the Brownian Corpus in the US National Corpus [11, 12]. The tagging systems are described in more details in [13, 14]. All these systems are mainly used for tagging the English language.

Currently there are more than ten electronic corpora for Turkic languages: the Kazakh language corpus [15–17]; the Tatar language corpus "Tugan tel" [18, 19], the Turkish language corpus [20]; the Crimean Tatar language corpus [21], Chuvash language [22], etc. One of the main components of these corpora is the system of morphological, syntactic and semantic tagging. Morphological tagging systems vary in some corpora.

There are various domain models based on ontology. For example, [23] considers the ontological model of hotel services in order to compare actual reviews with automatically generated ones. Some works [24, 25] present formal models for nouns in the Kazakh language and the use of semantic graphs to describe ontological models. The articles [26, 27] compare the ontological models on the example of the nouns in the Kazakh language with the other Turkic (Kyrgyz and Uzbek) languages, and [28] compares adjectives in the Kazakh and the Turkish languages.

3 Metalanguage for Kazakh Morphology

Table 1 presents the UniTurk tags on the example of the morphology concepts description for the Kazakh language. The first column of the table contains tag designations, the second - designations of the concepts as names in English, the third – tag names in Kazakh.

Table 1. Tags of the concepts of the Kazakh language morphology.

Tag	Name_English	Name_Kazakh
N	Noun	Зат есім
ANIM	Animate Noun	Жанды зат есім
INAM	Inanimate Noun	Жансыз зат есім
CMMN	Common Noun	Жалпы зат есім
PRPR	Proper Noun	Жалқы зат есім
CNCR	Concrete Noun	Деректі зат есім
ABST	Abstract Noun	Дерексіз зат есім
CASES_category	Cases	Септік жалғауы
NOM	Nominative case	Атау септік
GEN	Genitive case	Ілік септік
DIR	Direction- dative case	Барыс септік
ACC	Accusative case	Табыс септік
LOC	Locative case	Жатыс септік
ABL	Ablative case	Шығыс септік
INST	Instrumental case	Көмектес септік
NUMBER_category	Plural endings	Көптік жалғауы
SG	Singular	Жекеше
PL	Plural	Көпше
PERSON_category	Personal ending	Жіктік жалғауы
P1SG1	1 personal singular	1 жақ жекеше
P2SG1	2 personal singular	2 жақ жекеше (анайы)
P2SG.P1	2 personal singular formal	2 жақ жекеше (сыпайы)
P3SG	3 personal singular	3 жақ жекеше
P1PL1	1 personal plural	1 жақ көпше
P2PL1	2 personal plural	2 жақ көпше (анайы)
P2PL.P1	2 personal plural formal	2 жақ көпше (сыпайы)
P3PL	3 personal plural	3 жақ көпше
POSS_category	Possessive ending	Тәуелдік жалғауы
POSS.1SG	1 possessive singular	1 жақ жекеше

(*continued*)

Table 1. (*continued*)

Tag	Name_English	Name_Kazakh
POSS.2SG	2 possessive singular	2 жақ жекеше (анайы)
POSS.2SG.P	2 possessive singular formal	2 жақ жекеше (сыпайы)
POSS.3SG	3 possessive singular	3 жақ жекеше
POSS.1PL	1 possesive plural	1 жақ көпше
POSS.2PL	2 possesive plural	2 жақ көпше (анайы)
POSS.2PL.P	2 possesive plural formal	2 жақ көпше (сыпайы)
POSS.3PL	3 possesive plural	3 жақ көпше
Suffixes	Suffixes	Жұрнақ
WF	Word-formative	Сөз тудырушы жұрнақтар
NNWF	Suffixes that form the noun from nouns	Зат есімнен зат есім тудыратын жұрнақтар
VNWF	Suffixes that form the noun from verbs	Етістіктерден зат есім тудыратын жұрнақтар
AdjNWF	Suffixes that form the noun from adjectives	Сын есімнен зат есім тудыратын жұрнақтар
NumNWF	Suffixes that form the noun from numerals	Сан есімнен зат есім тудыратын жұрнақтар
WC	Word inflection	Сөз түрлендіруші жұрнақтар
NWC	Suffixes that change the meaning of nouns	Зат есімнің рең мәнін тудыратын жұрнақтары
HIP	Hipocoristic	Еркелету
DIM	Diminutive	Кішірейту
SIM	Similative	Келемеждеу, мысқылдау
HON	Honorific	Сыйлау, құрметтеу, үлкен тұту
PEJ	Pejorative	Менсінбеу, кемсіту
V	Verb	Етістік
Adj	Adjective	Сын есім
Num	Numeral	Сан есім
Pron	Pronoun	Есімдік
Adv	Adverb	Үстеу
FW	Function words	Шылау
Intrj	Interjection	Одағай
IW	Imitative words	Еліктеу сөздер

The text part consists not only of the name of the concepts, but also the definitions, questions, and examples in Kazakh language. As a result of the study, a tagging system for morphological concepts of the Kazakh language was obtained (Fig. 1).

Tag	Name_English	Name_Kazakh	definition	questions	example
N	Noun	Зат есім	заттың, құбылыстың, оқиғаның атын білдіреді	кім? не?	
CASES category	Case ending	Септік жалғауы	сөйлемдегі сөздерді бір-бірімен жалғасытырып, септестіріп тұратын қосымшалар		
NOM	Nominative case	Атау септік		кім? не?	бала
GEN	Genitive case	Ілік септік		кімнің? ненің?	баланың
DIR	Direction- dative	Барыс септік		кімге? неге?	балаға
ACC	Accusative(initial)	Табыс септік		кімді? нені?	баланы
LOC	Locative case	Жатыс		кімде? неде?	балада
ABL	Ablative case	Шығыс		кімнен?	баладан
INST	Instrumental case	Көмектес		кіммен? немен?	баламен
NUMBER category	Plural endings	Көптік жалғауы	нәрсенің көптігін білдіретін		
SG	Singular	жекеше		кім? не?	бала
PL	Plural	көпше		кімдер? нелер?	балалар
PERSON category	Personal ending	Жіктік жалғауы			
P1SG1	1 personal singular	1 жақ			Мен баламын
P2SG1	2 personal singular	2 жақ			Сен баласың
P2SG.P1	2 personal singular	2 жақ			Сіз баласыз
P3SG	3 personal singular	3 жақ			Ол бала
P1PL1	1 personal plural	1 жақ көпше			Біз баламыз
P2PL1	2 personal plural	2 жақ көпше			Сендер баласындар
P2PL.P1	2 personal plural formal	2 жақ көпше сыпайы			Сіздер баласыздар
P3PL	3 personal plural	3 жақ көпше			Олар балалар
POSS category	Possesive ending	Тәуелдік жалғауы	үш жақтың біріне белгілі бір заттың меншікті екенін білдіретін	кімім? нем? кімің? нең?	

Fig. 1. Fragment of a tagging system for morphological concepts of the Kazakh language.

4 Knowledgebase for Kazakh Morphology

4.1 Architecture

Morphology is a section of grammar, the main objects of which are words of the natural languages, their significant parts, and morphological features. The tasks of morphology, therefore, include the definition of a word as a special language object and a description of its internal structure [29]. According to this definition, the following concepts should be covered in the knowledge base architecture: *word, morpheme, parts of speech* and *morphological category*.

Word

A word is the central unit of a language.

Morpheme

In words, the minimum significant distinguished parts are morphemes. The main morpheme is the root, which bears the main meaning of the word, and the other morphemes are called affixes (suffix and ending).

1. Morpheme
   1.1 Root
   1.2 Affix
      1.2.1 Suffix
      1.2.2 Ending

Parts of Speech

Parts of speech are classes into which words are distributed according to their grammatical properties. There are nine parts of speech in the Kazakh language:

1. Noun
2. Adjective
3. Numeral
4. Pronoun
5. Verb
6. Adverb
7. Conjunction
8. Interjection
9. Imitative words

All parts of speech have certain properties, and all these properties are implemented in ontology.

Morphological Category

The morphological category is the affixal variations of words, which is an expression of the grammatical category of this word pattern. In the Kazakh language, there are different morphological categories, for example, for a noun (case category, number, person and possessiveness), adjective (degree of comparison) and verb (mood category, time, person, type, voice), etc.

1. The category of CASE
   1.1. Genitive case
   1.2. Direction-dative case
   1.3. Accusative case
   1.4. Ablative case
   1.5. Locative case
   1.6. Instrumental case
2. The category of NUMBER
3. The category of PERSON
4. The category of POSSESSIVE
5. Degrees of Comparison
6. The category of MOOD

7. The category of TENSE
8. The category of VOICE

All the above-mentioned concepts of the morphology of the Kazakh language were included into the applied ontology.

4.2 Implementation

Applied ontology "The morphology of the Kazakh language" consists of separate individuals, properties and classes, as well as interpretation functions defined on the ontology concepts or relations.

In this section, the ontology "Morphology of the Kazakh language" was implemented in accordance with the architecture in the Protégé environment. This environment is used to represent knowledge in the form of classes, individuals belonging to classes, and properties between them. All classes of morphological concepts of the Kazakh language are displayed in Fig. 2.

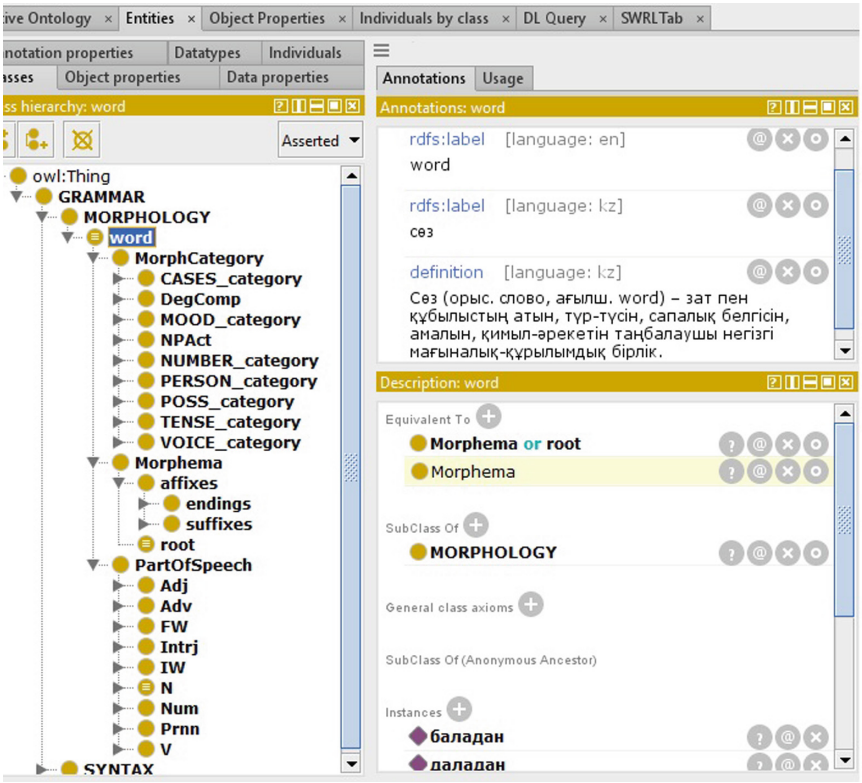

Fig. 2. Ontology "Morphology of the Kazakh language".

After the classes were created, the fields – properties – were written in them. Properties of objects define some relationship between two objects (classes, individuals). For example, the concept "Adj" (adjective) is characterized by the following properties (Fig. 3): the type can be either "qualitative adjective" (Qual) or "relative adjective" (Rel), in which the meanings of adjectives change. Therefore, we introduce the property "hasSemanticType" (has a semantic meaning), and describe it by an existential constraint (or limitation by a quantifier of existence):

$$\exists R.(A \cup B) \tag{2}$$

where $R = hasSemanticType$, $A = Qual$, $B = Rel$.

It can also specify that the adjective has the category of comparison degree:

$$\exists R.A \tag{3}$$

where $R = hasMorphCategory$, $A = DegComp$.

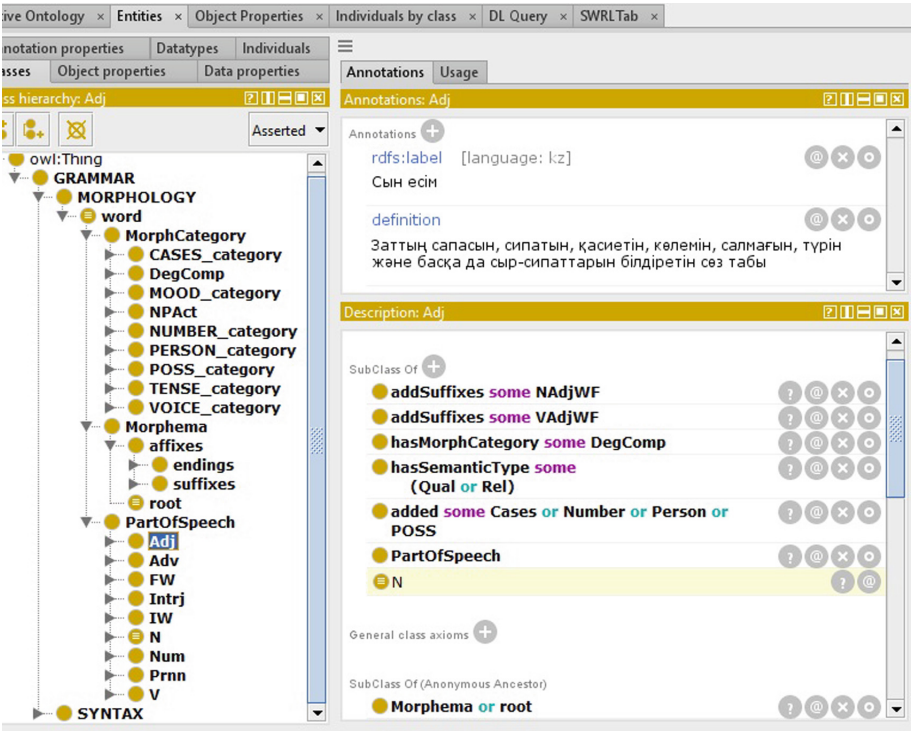

Fig. 3. Example of properties of adjectives in the Kazakh language.

The Fig. 4 shows some properties of nouns in the Kazakh language, which are defined by the following formulae:

$$\exists addSuffixes.(NNWF) \tag{4}$$

$$\exists addSuffixes.(VNWF) \tag{5}$$

$$\exists addSuffixes.(WC) \tag{6}$$

$$\exists added.(Number \cup POSS \cup Person \cup Cases) \tag{7}$$

$$\exists hasSemanticType.(ABST \cup CNCR) \tag{8}$$

$$\exists hasSemanticType.(ANIM \cup INAM) \tag{9}$$

$$\exists hasSemanticType.(CMMN \cup PRPR) \tag{10}$$

$$\exists hasMorphCategory.(CASES_category \cup NUMBER_category \cup PERSON_category \cup POSS_category) \tag{11}$$

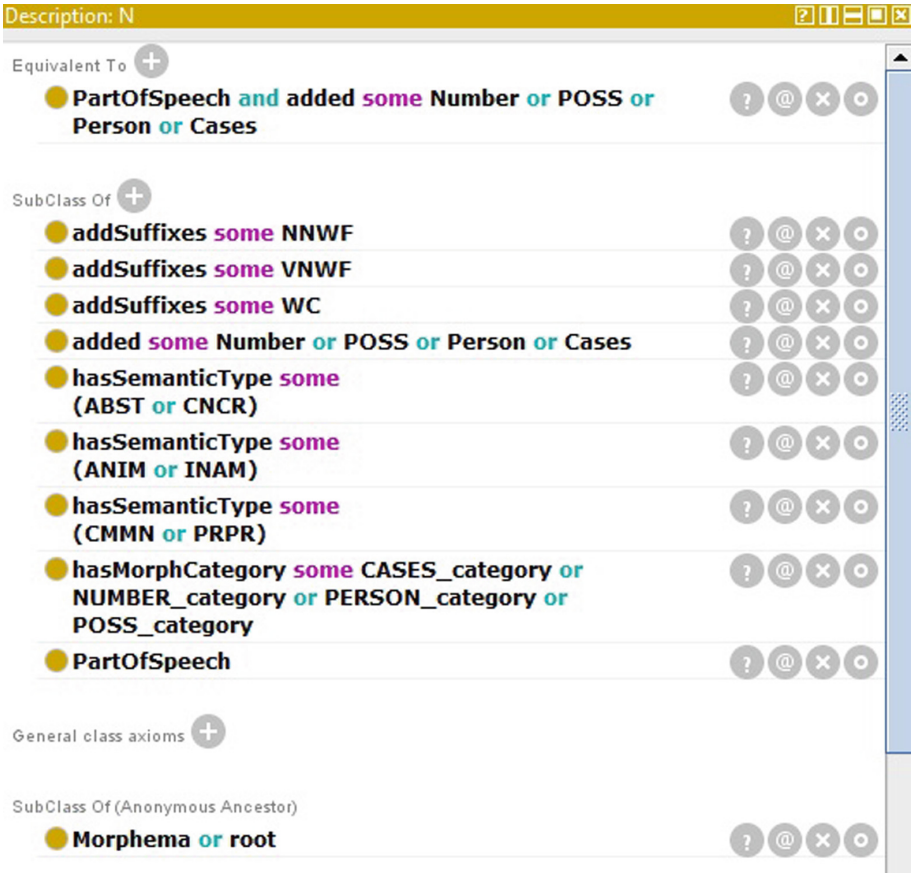

Fig. 4. Example of the properties of nouns in the Kazakh language.

There is also the considered concept of substantivization, which is a transition of other parts of speech (adjectives, verbs, participles, numerals) to the category of nouns, due to their ability to directly indicate the subject (which means it will answer the question “who?” or “what?”). For the noun, the necessary condition is to add to it the ending (Number, POSS, Person, Cases). In order to get the substantivization of other parts of speech into a noun, we will convert the necessary conditions into necessary and sufficient conditions (Fig. 5).

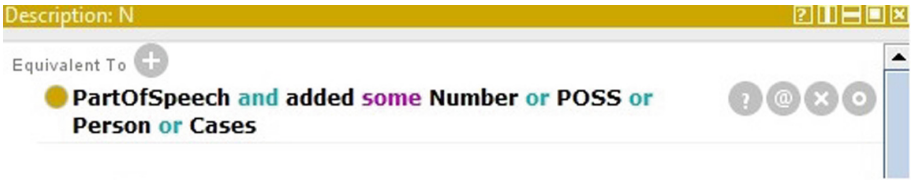

Fig. 5. Necessary and sufficient conditions for the noun.

When starting the reasoner, we obtain the substantivization of the adjective (Fig. 6), because in the process of word usage some adjectives lose their qualitative semantics and acquire subject meaning, i.e. substantivate, go into the category of nouns, thereby, replenishing vocabulary.

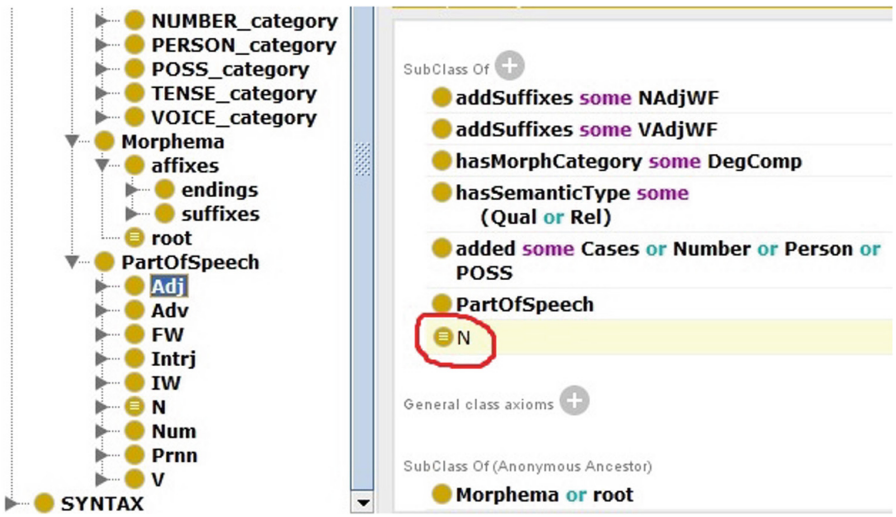

Fig. 6. Substantivization of adjectives.

For the morphological category NAbl (noun in the ablative case) there are necessary and sufficient conditions:

$$NAbl \equiv \exists hasRoot(root \cap (\forall hasPOS(N)) \cap \exists hasEnding(AblEnd)) \quad (12)$$

This allows defining the word "даладан - from outside", which is an individual of the concept "word" in the category NAbl. The Fig. 7 presents the description of the category NAbl and its implementation after launching the reasoner.

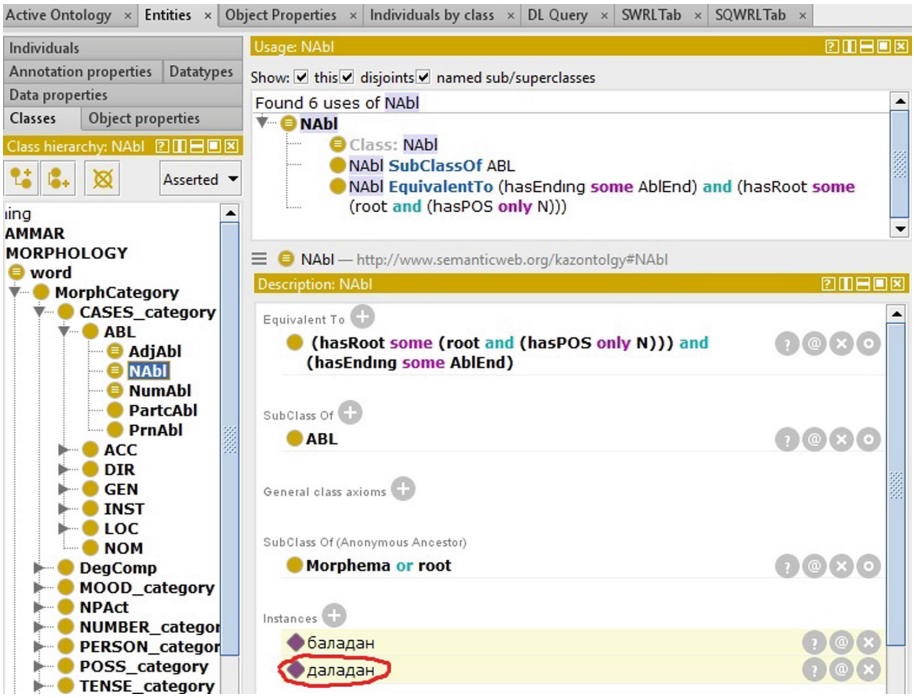

Fig. 7. Description of the NAbl category.

Thus, the created ontological model "Morphology of the Kazakh language" (Fig. 8) includes all the concepts and relations between them that are associated with the morphological features of the Kazakh language [30–32].

To formalize the morphological rules of natural languages, it is also possible to use alternative models, such as formal grammars, and languages of functional, logical, and production programming, and others.

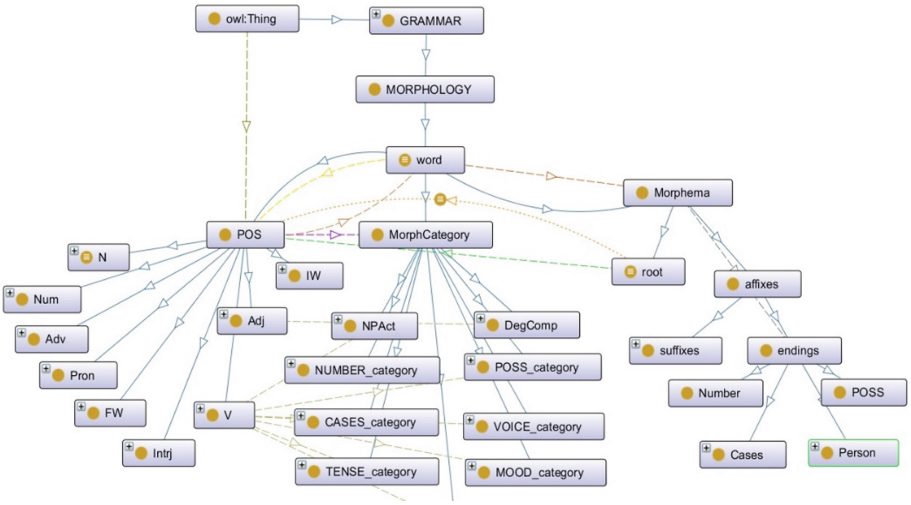

Fig. 8. Ontological model "Morphology of the Kazakh language".

5 Conclusion

Based on the study of metalanguages and methods for ontological modeling of subject areas, the metalanguage UniTurk was developed, it presented the notation of all the concepts of the Turkic languages morphology on the example of the Kazakh language and the ontology "Morphology of the Kazakh language" was built.

In the future, all syntactic concepts of the Turkic languages will be added to the UniTurk metalanguage and Grammar of the Turkic Languages will be created.

The linguistic resources created allow, on the one hand, to promote mutual understanding of terminology between the Turkic languages, and on the other hand, to become a multilingual knowledge base that will be used in multilingual search systems, machine translation between the Turkic languages, auto-referencing of the Turkic texts, and in information and training systems.

Acknowledgments. The work was supported by the grant financing for scientific and technical programs and projects by the Ministry of Science and Education of the Republic of Kazakhstan (Grant No. AP05132249, 2018–2020).

References

1. Bekmanova, G., et al.: A uniform morphological analyzer for the Kazakh and Turkish languages. In: Proceedings of the Sixth International Conference on Analysis of Images, Social Networks and Texts, Moscow, Russia, pp. 20–30 (2017)
2. Zhetkenbay, L., Sharipbay, A.A., Bekmanova, G.T., Kazhymukhan, D., Kamanur, U.: Comparison of the morphological rules of the Kazakh and Turkish languages. J. Math. Mech. Comput. Sci. **100**(4), 42–51 (2018)

3. Sharipbaev, A.A., Bekmanova, G.T., Buribayeva, A.K., Yergesh, B.Zh., Mukanova, A.S., Kaliyev, A.K.: Semantic neural network model of morphological rules of the agglutinative languages. In: Proceedings of the 6th International Conference on Soft Computing and Intelligent Systems, The 13th International Symposium on Advanced Intelligent Systems, Kobe, Japan, pp. 1094–1099 (2012)
4. Tsukanova N.I.: Ontological model for knowledge representation and organization, Moscow (2015)
5. Smekhun Y.A.: Ontologies in the knowledge based systems: possibilities of their application. Int. Res. J. https://doi.org/10.18454/IRJ.2016.47.086. Accessed 25 Feb 2019
6. Protégé. http://protege.stanford.edu. Accessed 10 Mar 2019
7. Gruber, T.R.: Toward principles for the design of ontologies used for knowledge sharing. Int. J. Hum Comput Stud. **43**(5–6), 907–928 (1995)
8. Zalevskaya A.A.: Introduction to Psycholinguistics: A Textbook. Moscow (2000)
9. Marcus, M.P., Santorini, B., Marcinkiewicz, M.A.: Building a large annotated corpus of English: the Penn Treebank. Comput. Ling. pp. 313–330 (1993)
10. Garside, R.: The CLAWS word-tagging System. In: Garside, R., Leech, G., Sampson, G. (eds.) The Computational Analysis of English: A Corpus-based Approach. Longman, London (1987)
11. The Open American National Corpus. http://www.anc.org. Accessed 10 Dec 2018
12. Ide, N.: The American national corpus: then, now, and tomorrow. In: Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus: Mustering Languages, Cascadilla Proceedings Project, Sommerville, MA (2008)
13. Jurafsky, D., Martin, J.H.: Speech and Language Processing: An Introduction to Natural Language Processing. Computational Linguistics and Speech Recognition. 2nd edn. Prentice-Hall (2009)
14. Indurkhya, N., Damerau, F.J.: Handbook of Natural Language Processing. 2nd edn. Chapman & Hall/CRC (2010)
15. Kazakh Language Corpus. http://kazcorpus.kz/. Accessed 10 Dec 2018
16. Makhambetov, O., Makazhanov, A., Yessenbayev, Zh., Matkarimov, B., Sabyrgaliyev, I., Sharafudinov, A.:. Assembling the Kazakh language corpus. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1022–1031 (2013)
17. Madiyeva, G.B., Umatova, ZhM: About Almaty Corpus of the Kazakh language KazNU messenger. "Phology" Series **5**(157), 99–103 (2015)
18. Tatar National corpus "Tugan tel". http://tugantel.tatar. Accessed 10 Dec 2018
19. Galieva, A., Khakimov, B., Gatiatullin A.: On the way to the relevant grammatical tagset for the tatar national corpus. In: 8th International Conference on Corpus Linguistics, Málaga, Spain, pp. 121–129 (2016)
20. Turkish National Corpus (TNC). http://www.tnc.org.tr/. Accessed 10 Dec 2018
21. Kubedinova, L., Gatiatullin, A.: Morphological tagging of crimean tatar electronic corpus. In: Proceedings of the international conference "Turkic languages processing", Kazan, Tatarstan, pp. 331–337 (2015)
22. Zheltov, P.: Morphological annotation system for the national corpus of the chuvash language. In: Proceedings of the international conference "Turkic languages processing", Kazan, Tatarstan, pp. 328–331 (2015)
23. Bekmanova, G., Sharipbay, A., Omarbekova, A., Yelibayeva G., Yergesh B.: Adequate assessment of the customers' actual reviews through comparing them with the fake ones. In: Proceedings of 2017 International Conference on Engineering & Technology (ICET 2017), Antalya, Turkey (2017). ISBN: 978-1-5386-1949-0

24. Yergesh, B., Mukanova, A., Sharipbay, A., Bekmanova, G., Razakhova, B.: Semantic hyper-graph based representation of nouns in the Kazakh language. Computacion y Sistemas **18**(3), 627–635 (2014)
25. Mukanova, A., Yergesh, B., Bekmanova, G., Razakhova, B., Sharipbay, A.: Formal models of nouns in the Kazakh language. Leonardo Electron. J. Practices Technol. **13**(25), 264–273 (2014)
26. Aripov, M., Sharipbay, A., Abdurakhmonova, N., Razakhova B.: Ontology of grammar rules as example of noun of Uzbek and Kazakh languages. In: Abstract of the VI International Conference "Modern Problems of Applied Mathematics and Information Technology - Al-Khorezmiy 2018", pp. 37–38, Tashkent, Uzbekistan (2018)
27. Sharipbay, A., Yergesh, B., Yelibayeva, G., Israilova, N., Bakasova, P., Zhetkenbay L.: Ontological models matching of nouns of Kazakh and Kyrgyz languages. In: Proceedings of the International Conference on Computer Processing of Turkic Languages, Tashkent, Uzbekistan, pp. 182–188 (2018)
28. Zhetkenbay, L., Sharipbay, A., Bekmanova, G., Kamanur, U.: Ontological modeling of morphological rules for the adjectives in Kazakh and Turkish languages. J. Theor. Appl. Inf. Technol. **91**(2), 257–263 (2016)
29. Momynova, B.K., Satkenova, ZhB: Morphology of the Kazakh language: a teaching aid. Almaty, Kazakhstan (2014)
30. Yskakov, A.: Modern Kazakh Language, 2 edn. Almaty, Kazakhstan (1991)
31. The Kazakh grammar. Phonetics, word formation, morphology, syntax. Astana, Kazakhstan (2002)
32. Institute of State language development: The Kazakh Language (Short grammar reference book). Almaty, Kazakhstan (2010)