A Uniform Morphological Analyzer for the Kazakh and Turkish Languages

Gulmira Bekmanova[1], Altynbek Sharipbay[1], Gulila Altenbek[2], Eşref Adalı[3], Lena Zhetkenbay[1], Unzila Kamanur[1], Altanbek Zulkhazhav[1]

[1] L.N. Gumilyov Eurasian National University, Astana Kazakhstan,
[2] College of Information Science and Engineering, Xinjiang University, P.R. China
[3] Istanbul Technical University, Istanbul, Turkey,
gulmira-r@yandex.ru, sharalt@mail.ru, jetlen_7@mail.ru unzila.88@mail.ru, altinbekpin@gmail.com, glaxd2014@163.com, esrefadali@gmail.com

Abstract. The Kazakh and Turkish languages belong to the group of the Turkic languages and have much in common. The detailed comparison of the ontologies on the example of the Kazakh and Turkish nouns allowed entering the analysis of morphological rules of these languages and the unified system of designations to create the uniform morphological analyzer based on the general algorithm of the morphological analysis.

Keywords: morphological analysis of the Kazakh and Turkish languages, ontology, analysis of morphological rules

1 Introduction

One of the methods to reduce the semantic barrier between the human and the computer is searching new methods of a natural language processing. Nowadays it is obvious that in order to implement the human-computer interaction in a natural language and to create a linguistic support of the information processes the study of the language itself is required. Besides the resources consumed could be decreased due to formalization of language rules providing the storage of information in procedural but not in declarative form. For the Kazakh and the Turkish languages which morphological regularities are quite well yielded to formalization, it would produce an excellent result.

All language levels are characterized by existence of basic elements. A language studying can take place from two positions – the analysis and synthesis because the revealed rules of synthesis can assist to carry out the analysis and vice versa. In this case the Kazakh and Turkish languages are studied from both positions the analysis and synthesis. This very integrated approach allows to study in details all regularities and to reveal such nuances which when using only of one of approaches would remain outside our attention. For researching and the maximum formalizing of each language subsystem it is necessary to create the program tools implementing the studying process by identifying and verifying

the analysis and synthesis rules. There-with it will greatly automate the research process and a researcher does not need to accumulate and collect information. And the labor intensity is very low.

The morphology modeling is related to all applications such as natural language and tasks processing and includes information search, moods analysis, spelling correction, detection of the generated texts, parts of speech marking and entity extraction. The morphology is used in linguistics to refer to the study of structure and formation of words. The Agglutinative languages (agglutinare from Latin means to stick together) are languages which morphological system is characterized by agglutination ("pasting") of various formants. As a formant either prefixes or suffixes act and each of them makes its own sense.

As the Kazakh and Turkish languages belong to the group of Turkic languages and the languages of this group can be classified as agglutinative languages. These languages are full of word forms (inflections). Inflections are formed by addition of suffixes. The suffixes are attached in the strict sequence and the resulting new words can belong to the other part of speech. The possessive form in Kazakh is similar to a possessive form in English [1, 2]. Plenty of researches covering formalization of morphological rules and morphological analyses of [3-6] the Turkic languages are avail-able. The first morphological analyzer of Kazakh was developed in 2009 and based on the procedural method. The procedural method implies the preliminary systematization of morphological knowledge about a natural language and development of morphological information assignment algorithms to a separate word form [7, 8].The procedural morphological analyzer of Kazakh consisted of the following stages: marking the stem in the current word form, its identification, assigning to a word form the corresponding list of morphological information. The disadvantage of this method is high labor intensity while compiling the dictionaries of compatibility. This challenge is difficult to be settled and cannot be automated completely for languages which are characterized by a large number of counterexamples. The implementation of this method occupies considerably smaller memory size, but at the same time the morphological analysis period due to splitting a word form into components and applying the procedures of compatibility increases [8]. The second version of the morphological analyzer was developed in 2012 and based on the formal morphological rules [9]. Later versions were based on using the ontological models and the hyper graphs [10-13]. The other research groups developed their own morphological analyzers [15-16].

The works on creation of the morphological analysis for the Turkish language are carried out for a long period of time and presented in papers [17-21]. In this paper the results received in [17] were used. The peculiarity of this morphological analyzer is the methodology for carrying out the analysis. The Turkish words with affixation were used without any lexicon. This morphological ana-

lyzer is completely based on the rules and implies using only the dictionary of counterexamples. The analyzer is based on the final automatic model.

2 The generalized ontologic models of parts of speech of the Kazakh and Turkish languages

Ontology is a powerful and widely used tool for modelling relationships between objects which belong to the different subject area. Ontology should be classified based on the degree of dependence on the task or application area, ontological model for knowledge representation and expression as well as other criteria [22].

We used the ontology editor Protg (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.

The morphological features of initial forms of nouns (N) are as follows. A noun can be either animate (anim) or inanimate (inanim); this feature determines the trajectory of the inflection of a noun. Nouns in the Kazakh language can be conjugated (pers_end) and vary for case (cases) and number (number), as well as have a possessive form (poss_end) [10].

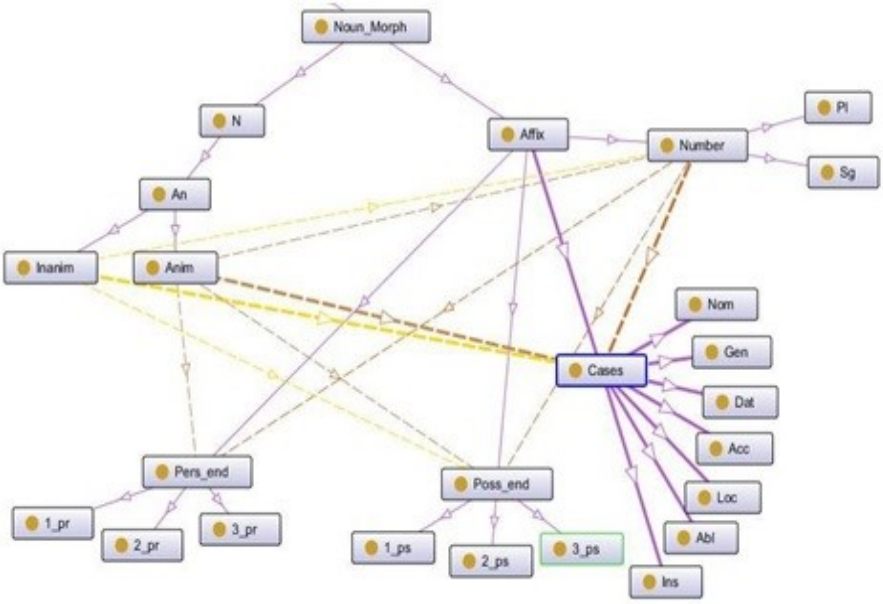

Fig.1: Ontological model of the Kazakh noun [10]

Figure 1 shows the ontological model of the Kazakh noun with its morphological features. Concepts and relationships used in this ontological model are explained

in Table 1.

Table 1. Concepts and relationships

Notation	Description	Notation	Description
N	Noun	2 pr	2 personal
Part of speech		3 pr	3 personal
Item	Item	Poss_end	Possessive endings
Anim	Animate	1 ps	1 personal
Sign of animacy	v16	2 personal	
Inanim	Inanimate	3 ps	3 personal
Sign of inanimacy		Number	Number
Cases	Cases	Pl	Plural
Nom	Nominative case	Sg	Singular
Gen	Genitive case	is_a	
Dat	Direction- dative case	Denotes	
Acc	Accusative		
e3, e4	has_feature		
Loc	Locative case	Has	
Abl	Ablative case	Devided	
Ins	Instrumental case	Change	
Pers_end	Personal endings	Add	
1 pr	1 personal		

The ontology model of the Kazakh parts of speech allows us to completely de-scribe the morphological rules and their relationships. On the basis of this ontological model we developed generalized ontological models of the Kazakh and Turkish language parts of speech. The developed ontological model of nouns of the Kazakh language in Protege environment is displayed in the Figure 2, and the ontological model of nouns of the Turkish language is shown below in the Figure 3.

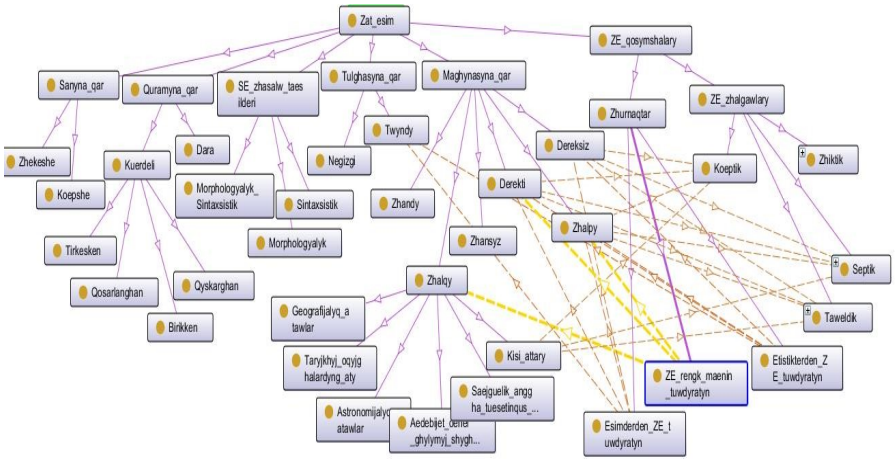

Fig.2: The ontological models of nouns for the Kazakh language

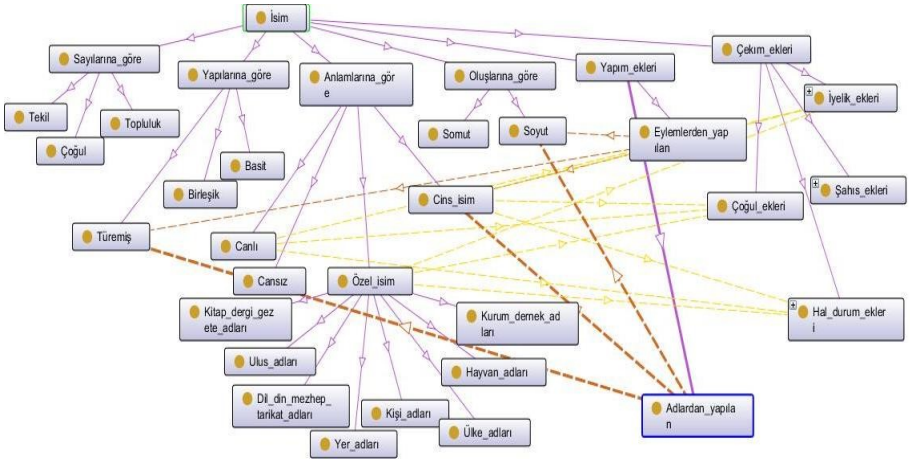

Fig.3: The ontological models of nouns for the Turkish language

In this way the comparative ontological models of noun for machine translation system include all the categories of morphological features, for instance, noun is divided as stem and complex according to the structure of noun in the Kazakh language whereas in the Turkish language there is not such division, furthermore, a noun can be common, proper, concrete, abstract, animated, inanimate according to meaning in the Kazakh language, while in the Turkish language a noun can be common, proper, animated, inanimated. In both languages the divisions of affixation are similar, e.g., the forms of cases, number, possessives and conjugations. There are seven cases in Kazakh whereas in Turkish there are five.

3 The uniform morphological analyzer for the Kazakh and Turkish languages

The comparison of the ontological models allowed creating the general symbol system of morphological markers which are used in morphological analyzer.

Table 2. - The comparison of morphological markers of the Kazakh and Turkish languages nouns

N	Abbreviation	Name in English	Name in Kazakh	Name in Turkish	Unified Tag
1	+Noun	Noun	Zat esim	İsim	Noun
2	+A1sg	Personal 1 singular	Zhiktik 1 zhaq, zhekeshe	1. Tekil Şahıs Uyum Özelliği	PERS.1SG
3	+A2sg	Personal 2 singular	Zhiktik 2 zhaq, zhekeshe	2. Tekil Şahıs Uyum Özelliği	PERS.2SG
4					PERS.2SG.POL
5	+A3sg	Personal 3 singular	Zhiktik 3 zhaq, zhekeshe	3. Tekil Şahıs Uyum Özelliği	PERS.3SG
6	+A1pl	Personal 1 plural	Zhiktik 1 zhaq, koepshe	1. Çoğul Şahıs Uyum Özelliği	PERS.1PL
7	+A2pl	Personal 2 plural	Zhiktik 2 zhaq, koepshe	2. Çoğul Şahıs Uyum Özelliği	PERS.2PL
8					PERS.2PL.POL
9	+A3pl	Personal 3 plural	Zhiktik 3 zhaq, koepshe	3. Çoğul Şahıs Uyum Özelliği	PERS.3PL
10	+P1sg	Possessive 1 singular	Zhiktik 1 zhaq, Taweldik 1 zhaq, zhekeshe	1. Tekil Şahıs Iyelik Eki	POSS.1SG
11	+P2sg	Possessive 2 singular	Taweldik 2 zhaq, zhekeshe	2. Tekil Şahıs Iyelik Eki	POSS.2SG
12	+P2sgpol	Possessive 2 singular (formal)	Taweldik 2 zhaq, zhekeshe, resmi tueri		POSS.2SG.POL

13	+P3sg	Possessive 3 singular	Taweldik 3 zhaq, zhekeshe	3. Tekil Şahıs Iyelik Eki	POSS.3SG
14	+P1pl	Possessive 1 plural	Taweldik 1 zhaq, koepshe	1. Çoğul Şahıs Iyelik Eki	POSS.1PL
15	+P2pl	Possessive 2 plural	Taweldik 2 zhaq, koepshe	2. Çoğul Şahıs Iyelik Eki	POSS.2PL
16	+P2plpol	Possessive 2 plural (formal)	Taweldik 2 zhaq, koepshe, resmi tueri		POSS.2PL.POL
17	+P3pl	Possessive 3 plural	Taweldik 3 zhaq, koepshe	3. Çoğul Iyelik Eki	POSS.3PL
18	+Pnon	Non Possessive	Taweldenbegen	Belirsiz İyelik	NON. POSS
19	+Nom	Nominative	Atau	Yalın Durum	NOM
20	+Acc	Accusative (whom?)	Tabys	Belirtme Durumu	ACC
21	+Dat	Dative	Barys	Yönelme Durumu	DAT
22	+Abl	Ablative	Shyghys	Çıkma Durumu	ABL
23	+Loc	Locative (where?)	Zhatys	Kalma Durumu	LOC
24	+Gen	Genitive (whose?)	Ilik	Tamlayan Durumu	GEN
25	+Ins	Instrumental	Koemektes	Aracılık Durumu	INS
26	+Pos	+Positive	Bolymdy	Olumlu	POSIT
27	+Neg	+Negative	Bolymsyz	Olumsuz	NEGAT

For example, the line 4 of the above-mentioned table does not have any meanings in the Kazakh and Turkish columns, there is no analogue in English, but preserved name means that for the other Turkic languages such morphological marker for noun exists. In the lines 12 and 16 the blank value in the Turkish language means that this morphological marker exists only for the Kazakh language.

Metalanguage is one of key concepts of system of the description of an object of science and is defined as artificial language of "the second order" in relation to which natural human language acts as "language object", that is as a subject of a linguistic research. In our case a natural language are the Kazakh and Turkish

languages enter-ing into the Turkic group of languages.

The unified symbol system (UNIFIED TAG) was developed based on the idea of creating unified metalanguage for Turkic Languages.

Firstly, the idea of creating metalanguage was proposed at the 1st so-called International Conference on Computer processing of the Turkic Languages (TurkLang-2013) which was held in Astana on 3-4 October, 2013. A group of famous professors of technical sciences A.A. Sharipbay (Astana, Kazakhstan), D.SH. Suleimenov (Kazan, Tatarstan, Russia), Eşref Adalı(Istanbul, Turkey) is working on the creation of metalanguage.

At the UniTurk scientific-practical seminar of the conference the discussion of problems related to the unification of grammatical categories for the corpuses of the Turkic languages raised a great interest and held successfully.

For computerizing the Kazakh language it is very important step to research the computational linguistics of the other Turkic-speaking countries. From this point studying the structures of agglutinative languages that are similar to Kazakh and make comparisons between them leads to a successful computer transforming of all languages belonging to the Turkic languages group. We are very confident that it will bring great success in development of the Kazakh language computerizing.

Our goal is to use correctly these similarities and differences in the language automating direction. While entering to a computer the similarities between languages help to solve the unsolved problems in one language by supplementing the achievements of another language, moreover, studying the differences of languages according to its features in cooperation gives us an opportunity to implement a method in one language which didn't give any results in another language. The analysis made revealed that the Kazakh and Turkish languages have much in common. The Table 3 shows the comparison of the rules for a noun window in the Kazakh and Turkish languages.

Table 3. - Example of inflection a noun window

English	Kazakh	Turkish
Case endings of Noun (singular form)		
window	tereze: tereze+Noun+A3sg+Pnon+Nom	pencere: pencere+Noun+A3sg+Pnon+Nom
window 's	terezening: tereze+Noun+A3sg+Pnon+Gen	pencerenin: pencere+Noun+A3sg+Pnon+Gen

to window	terezege: tereze+Noun+A3sg+Pnon+Dat	pencereye: pencere+Noun+A3sg+Pnon+Dat
window	terezeni: tereze+Noun+A3sg+Pnon+ Acc	pencereyi: pencere+Noun+A3sg+Pnon+ Acc
window	terezede: tereze+Noun+A3sg+Pnon+ Loc	pencerede: pencere+Noun+A3sg+Pnon+ Loc
from window	terezeden: tereze+Noun+A3sg+Pnon+ Abl	pencereden: pencere+Noun+A3sg +Pnon+ Abl
with window	terezemen: tereze+Noun+A3sg+Pnon+Ins terezemen: tereze+Noun+A3sg+P1sg +Ins	pencerele: pencere+Noun+A3sg+ Pnon+ Ins pencerele: pencere+Noun+A3sg+P1sg+ Ins
Case endings of Noun (plural form)		
windows	terezeler: tereze+Noun+A3pl+Pnon+Nom	pencereler: pencere+Noun+A3pl+Pnon+Nom
windows'	terezelerding: tereze+Noun+A3pl +Pnon+Gen	pencerelerin: pencere+Noun+A3pl +Pnon+Gen
to windows	terezelerge: tereze+Noun+A3pl+Pnon+Dat	pencerelere: pencere+Noun+A3pl +Pnon+Dat
windows	terezelerdi: tereze+Noun+A3pl+Pnon+ Acc	pencereleri: pencere+Noun+A3pl +Pnon+ Acc
windows	terezelerde: tereze+Noun+A3pl+Pnon+ Loc	pencerelerde: pencere+Noun+A3pl +Pnon+ Loc
from windows	terezelerden: tereze+Noun+A3pl+Pnon+ Abl	pencerelerden: pencere+Noun+A3pl +Pnon+ Abl
with windows	terezelermen: tereze+Noun+A3pl +Pnon+Ins	pencerelerle: pencere+Noun+A3pl +Pnon+ Ins

The record of morphological rules in the unified form allowed to create the uniform rule-based algorithm of morphological analysis for the Kazakh and Turkish languages in the papers [9, 10, 17].

4 Conclusion

In the present scientific paper the morphological features of the Kazakh and Turkish languages are analyzed. The ontologies comparison is made, the uniform symbol system of morphological features is developed and the morphological rules of the Kazakh and Turkish languages are written over via new symbol system. The unified morphological analyzer is developed based on the general

morphological analysis algorithm.

In the future it is supposed to create the unified metalanguage of the Turkic languages that will allow reaching the new level the Turkic languages processing.

References

1. Batayeva, Z. (2012). Colloquial Kazakh, Routledge
2. Kazakh grammar. (2002). Phonetics, word formation, morphology, syntax (in Kazakh). Astana
3. Sharipbayev A. A., Bekmanova G. T.: The synthesis of word forms of Turkic language using semantic neural networks. Modern problems of applied mathematics and information technologies: abstracts Al Khorezmy, pp.145 (2009)
4. Sharipbayev, A.A., Bekmanova, G. T.: The building of logical semantics of the Kazakh words. The materials of the all-Russian conference with International participations Knowledge-Ontology-Theory (ZONT-09), Novosibirsk, pp. 246-249 (2009)
5. Tantuğ, A. C., Adalı, E., Oflazer, K.: Computer Analysis of the Turkmen Language Morphology. In: Salakoski T., Ginter F., Pyysalo S., Pahikkala T. (eds) Advances in Natural Language Processing. Lecture Notes in Computer Science, vol 4139, pp. 186 193, Springer, Berlin, Heidelberg (2006)
6. Orhun, M., Tantuğ, A. C., Adalı, E.: Rule Based Analysis of the Uyghur Nouns. Proceedings of the International Conference on Asian Language Processing (IALP). Chiang Mai, Thailand, 19 (1): pp. 33-43 (2008)
7. Bekmanova, G. T.: Some approaches to the problems of automatic inflection and morpholog-ical analysis in the Kazakh language. The newsletter of D. Serikbayev East Kazakhstan state technical university, Ust-Kamenogorsk, pp. 192-197 (2009)
8. Dobrushina, E. P., Savina, G. B., Gelbukh, A. G.: The system of an accurate morphological analysis and synthesis. The software of new information technology, Kalinin (1989)
9. Sharipbayev, A., Bekmanova, G., Mukanova, ., Buribayeva, A., Yergesh, B., Kaliyev, A.: Semantic neural network model of morphological rules of the agglutinative languages. The 6th International Conference on Soft Computing and Intelligent Systems The 13th International Symposium on Advanced Intelligent Systems, Kobe, Japan, 20-24 November 2012, pp. 1094-1099
10. Yergesh, B., Mukanova, A., Bekmanova, G., Sharipbay, A., Razakhova, B.: Semantic hy-per-graph based representation of nouns in the Kazakh language. Computacion y Sistemas; Volume 18, Issue 3, 1 July 2014, pp. 627-635
11. Mukanova, A., Yergesh, B., Bekmanova G., Razakhova, B., Sharipbay, A.: Formal models of nouns in the Kazakh language. Leonardo Electronic Journal of Practices and Technologies; Is-sue 25 (July-December), 2014 (13), pp. 264-273
12. Zetkenbay, L., Sharipbay, A., Bekmanova, G., Kamanur, U.: Ontological modeling of morphological rules for the adjectives in Kazakh and Turkish languages. Journal of Theoretical and Applied Information Technology, Vol. 91. No.2, pp. 257-263 (2016)
13. Kamanur U., Sharipbay A., Altenbek G., Bekmanova G., Zhetkenbay L.: Investigation and Use of Methods for Defining the Extends of Similarity of Kazakh Language Sentences. In: Sun M., Huang X., Lin H., Liu Z., Liu Y. (eds) Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. CCL 2016, NLP-NABD 2016. LNCS, vol 10035. Springer, Cham (2016)

14. Tukeyev, U., Zhumanov, Zh., Rakhimova, D., Kartbayev, A.: Combinational Circuits Model of Kazakh and Russian Languages Morphology. Abstracts of International Conference Computational and Informational Technologies in Science, Engineering and Education, pp. 241-242. Al-Farabi KazNU Press, Almaty (2015)
15. Kessikbayeva, G., Cicekli, I.: Rule-Based Morphological Analyzer of Kazakh Language. Linguistics and Literature Studies 4(1): pp. 96-104 (2016)
16. Makhambetov, O., Makazhanov, A., Sabyrgaliyev, I., Yessenbayev, Z. Data-Driven Morphological Analysis and Disambiguation for Kazakh. In: Gelbukh A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2015.LNCS, vol 9041, pp. 151-163. Springer, Cham (2015)
17. Eryğit, G., Adalı, E.: An affix stripping morphological analyzer for turkish. In Proceedings of the IASTED International Conference on Artificial Intelligence and Applications, Innsbruck, Austria, pp. 299304 (2004)
18. Eryğit, G., Adalı, E.: Synthetic Turkish Word Root Generation. Proceedings of the Turkish Artificial Intelligence and Neural Networks, TAINN, Canak-kale,Turkey (2003)
19. Akın, A.A., Akın, M.D.: Zemberek, an open source nlp framework for Turkic languages. Available at http://zemberek.googlecode.com
20. Hakkani-Tür, D. Z., Oflazer, K., Tür, G.: Statistical morphological disambiguation for agglutinative languages. In Proceedings of COLING. ICCL (2000)
21. Hankamer, J. Finite State Morphology and Left to Right Phonology. Proceedings of the West Coast Conference on Formal Linguistics 5. Stanford University (1986)
22. Gruber, T.R.: Toward Principles for the Design of Ontologies Used for Knowledge Sharing. International Journal Human-Computer Studies Vol. 43, Issues 5-6, 907928 (1995)