CEUR Workshop Proceedings 2020 vol.2780, pages 226-243

About turkic morpheme portal

Gatiatullin A., Suleymanov D., Prokopyev N., Khakimov B.
Kazan Federal University, 420008, Kremlevskaya 18, Kazan, Russia

Abstract

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). Doing scientific research in fields of turkology and agglutinative languages typology requires software that takes into account the structural and functional features of languages in question. This paper presents a description of the Turkic Morpheme Portal, in the development of which an integrated approach is used for the development of computer linguistic models and technologies for Turkic languages processing. This portal was created on the basis of the structural-parametric functional model of the Turkic morpheme and contains special linguistic databases that describe the categories of Turkic languages at different levels: morphological, syntactic, and semantic. The problems of developing complex multilingual linguistic models for low-resource languages and their software implementation are considered. The prospects of using the created portal as a base for the development of linguistic software, as well as an information and reference system, including a multilingual thesaurus, and as a platform for communication of specialists are given.

Keywords

Linguistic resource, Morphology, Multilingual model, Ontology, Thesaurus, Turkology

References

[1] Guzev, V.G., Pyotrovski R.G., Sherbak A.M.: O sozdanii mashinnogo fonda tyurkskikh yazykov [About creation of machine fund for Turkic languages]. Sovetskaya tyurkologiya [Soviet turkology], 2, 92-101 (1988).

[2] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. In: ICLR: Proceedings of the International Conference on Learning Representations Workshop Track, 1301-3781 (2013).

[3] Fellbaum, C.: WordNet: an Electronic Lexical Database. MIT Press, Cambridge, MA (1998).

[4] «Tugan Tel» Tatar National Corpus, http://tugantel.tatar, last accessed 2020/10/15.

[5] Turkish National Corpus (TNC), https://www.tnc.org.tr, last accessed 2020/10/15.

[6] Bashkir poetical corpus, http://web-corpora.net/bashcorpus/search/, last accessed 2020/10/15.

[7] Turklang Electronic Corpora, http://www.turklang.net/en/resources-for-turkic-languages/, last accessed 2020/10/15.

[8] Global Lexicostatistical Database, http://starling.rinet.ru/new100/mainr.htm, last accessed 2020/10/15.

[9] Maps for Turkic Languages, http://turk.polycorpora.org, last accessed 2020/10/15.

[10] Navigli, R., Ponzetto, S. P.: BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. Artificial Intelligence, 193, 217-250. Elsevier (2012).

[11] Palmer, M.: SemLink: Linking PropBank, VN and FrameNet. In: Proceedings of the Generative Lexicon Conference, GenLex-09, 13-17 (2009).

[12] Plungyan, V.A.: Obshchaya morfologiya: Vvedenie v problematiku: Uchebnoe posobie. [General morphology: Problem introduction: Schoolbook]. Editorial URSS, Moscow (2000).

[13] Gorodetsky, B.Y.: K probleme semanticheskoy tipologii [Onto semantic typology problem]. Moscow University Press, Moscow (1969).

[14] Suleymanov, D.S.: Sistemy i informatsionnye tekhnologii obrabotki estestvennoyazykovykh tekstov na osnove pragmaticheski-orientirovannykh lingvisticheskikh modeley [Systems and information tehcnologies of natural language processing on basis of pragmatically-oriented linguistic models]. Doctorate thesis on technical sciences, Kazan State University, Kazan (2000).

[15] Suleymanov, D.S., Gatiatullin, A.R.: Strukturno-funktsionalnaya kompyuternaya model tatarskikh morfem [Structural and functional computer model of Tatar morphemes]. FEN Tatarstan Academy of Sciences, Kazan (2003).