DEBRECENI EGYETEM FOLIA URALICA DEBRECENIENSIA 28.
FINNUGOR NYELVTUDOMÁNYI TANSZÉK DEBRECEN, 2021

# Enhancing the Erzya-Moksha dictionary automatically with link prediction

**Khalid ALNAJJAR – Jack RUETER –**
**Niko PARTANEN – Mika HÄMÄLÄINEN**

University of Helsinki
khalid.alnajjar@helsinki.fi, jack.rueter@helsinki.fi,
niko.partanen@helsinki.fi, mika.hamalainen@helsinki.fi

## Introduction

Dictionary and vocabulary work with the Mordvin languages date back to the late 17th century with the collections by N. Witsen (*Noord en oost Tartarye etc.* Amsterdam, 1692) (cf. Feoktistov 1976: 10–15; Maticsák 2017: 43–61; Mikola 1975). Specifically Erzya and Moksha vocabularies appear in research work beginning in the second half of the 19th century (cf. Ahlqvist 1861, Moksha; Wiedemann 1865, Erzya) as bilingual dictionaries with German as the target dictionary. During the 20th century the source languages, Erzya and Moksha, appear in several bilingual dictionaries with Russian as the target language with separate Russian-to-Erzya and Russian-to-Moksha dictionaries to complement them. It is worth mentioning the considerable amount of research done by Keresztes (1990, 1999, 2011, 2013).

The 1990s also sees a proliferation of dictionary activity both institutional and individual. This decade is marked by the publication of extensive Erzya-Russian (1993) and Moksha-Russian (1998) bilingual dictionaries as well as foreign Erzya-Finnish (1996), Moksha-Finnish (1998), Erzya-Hungarian (1998), and a bounty of small dictionaries of specific focus. This is also when a completed Dictionary of the Mordvin Dialects [Mordwinisches Wörterbuch] based on the Paasonen research collections appears (1990, 1992, 1994, 1996).

The first decade of the new millennium sees new dimensions in dictionary writing. There is the continuation of bilingual dictionary writing, observed in Erzya-German (2002), Moksha-Hungarian (2003), on the one hand, but there is also the monolingual dictionary of the Erzya language (2002), and a trilin-

DOI: 10.52401/fud/2021/01

gual dictionary Moksha-Erzya-Russian (2004), on the other. Additions and updates to Erzya and Moksha dictionary work continue to this day both in print and online, e.g. an Estonian-Erzya dictionary appeared online in 2019 (Soosaar – Erina 2019) and multilingual dictionaries are available for both Mordvin languages online (Hämäläinen – Rueter 2018). (An overview of the Mordvin dictionary literature: Maticsák 2014.)

The search for an Erzya-to-Moksha dictionary is still ongoing. One requirement for a good bilingual dictionary is the existence of monolingual dictionaries, such as the one by the Erzya writer Kuz′ma G. Abramov (2002), which is actually the first one of its kind for a non-majority Finno-Ugric language (other than Hungarian, Finnish and Estonian), and has only one counterpart in the Komi monolingual dictionary (Karmanova – Jakubiv 2016).

Although trilingual dictionaries of Erzya, Moksha and Russian (or other target languages) are a good start for arriving at Erzya-to-Moksha dictionaries, there are certain issues to be dealt with first. The Paasonen Mordvin Dialect Dictionary, which contains nearly 7000 Mordvin roots but aligns less than 50% of the roots as mutual to Moksha and Erzya (Rueter 2016, 2020), is, in fact, an exceptional work of cognate alignment for continued studies in all Mordvin idioms. This cognate alignment can, to some extent, also be observed in the trilingual dictionaries of 2004 (Moksha-Erzya-Russian with at least 3,610 translation pairs) and 2011 (Russian-Moksha-Erzya with parallel Moksha and Erzya translations for at least 10,123 Russian words). In the former (Poljakov 2004) there is an adherence to single-word lemmas in the Erzya and Moksha, while in the latter (*Русско-мокшанско-эрзянский словарь* [Russko-mokšansko-èrzjanskij slovar] 2011) the single-word strategy only applies to the Russian source language. Both books were intended to demonstrate the coherence of the two Mordvin literary languages, although the ordered alignment of Mordvin cognates does not always follow semantic or frequency criteria found in literature and everyday usage. Hence there is still a need for a translation dictionary between Erzya and Moksha, which, on the one hand, is based on descriptions of the Erzya and Moksha languages as they are actually used, and also affords ample contextual examples for both languages.

In 2019 work was begun with an online dictionary editing platform Ve′rdd (Alnajjar et al. 2020). The goal of this undertaking has been to provide uniform representation and access to multilingual data, where the relations between entries and their translations have to be complexly mapped. We also recognized the importance of collaborative editing online, as many earlier projects had edited lexical information in XML, and essentially one dictionary at the time. We understand multilingual dictionaries as a complex network of meanings and translations, and the only way to describe this adequately is within a platform where we can build connections between these different resources.

In this paper, we describe how we have used the new features that we have built into Ve′rdd to automatically extend the existing, digital Erzya-Moksha dictionary. We go through a sample of the results to validate the efficacy of our model.

## Related work

There has been a multitude of attempts to predict novel translations in bilingual and multilingual dictionaries. Here, we describe some of the most relevant work to our work. It is to be noted that computational research conducted on endangered languages is very different from similar research conducted on other, so-called „low-resourced" languages (Hämäläinen 2021). There have been several related approaches to extending semantic knowledge bases (Pasini – Navigli 2017; Gesese et al. 2020), however, their research questions are very far from our goals in this paper.

Lam – Kalita (2013) have proposed a method for reversing bidirectional dictionaries (e.g., reversing Hindi-English to English-Hindi). Their approach depends heavily on WordNet[1] being available for at least one of the languages in question. and uses the similarities between the words and their synonyms, hyponyms and hypernyms in WordNet to estimate the quality of the reverse translations. They have tested the method by reversing resource-poor and endangered language dictionaries (e. g. Karbi, Hindi and Assamese) to have English as the source language instead of the target language.

Lam et al. (2015) elaborated a way for authoring novel dictionaries for resource-poor languages. In their research paper, a dictionary of a low-resource language to a resource-rich language with a high-quality WordNet is needed. To translate a word from the source language to a new language, their method uses links between the English WordNet and existing multiple intermediate WordNets of other languages such as Finnish and Japanese to highlight the relevant words in the WordNets. Thereafter, each of these words are translated to the destination language in question using some of the existing machine translation systems such as Google Translate. The higher the agreement between multiple machine translation systems, the higher the score given to the translation.

[1] https://wordnet.princeton.edu/

An approach based on constraints for inducting new dictionaries for low-resourced languages within the same language family has been proposed by Wushouer et al. (2015). In their approach, a graph is built from two bilingual dictionaries (i.e. A-B and B-C, where B is the intermediate language), and new potential translation links are examined by treating the problem as conjunctive normal form (CNF) and using WPMaxSAT solver to identify the new translations.

A graph-based method for combining multiple Wiktionaries and inferring new translations using graph-based probabilistic inference measured by random walks was proposed by Soderland et al. (2009). The goal of their work is to construct a huge dictionary covering the well-resourced languages (e.g., English, French, Spanish, etc.) and suggest new dictionary translations; nonetheless, their work does not address endangered or resource-poor languages.

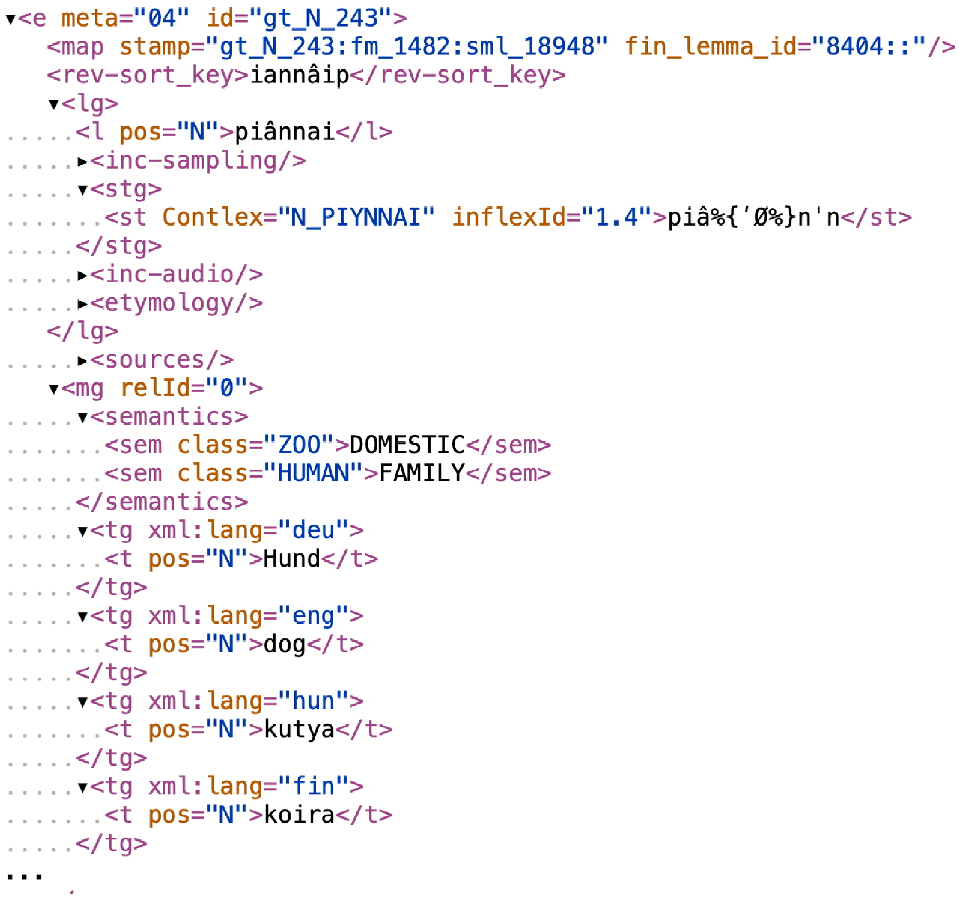
Donandt et al. (2017) have trained a Support Vector Machine (SVM) model to predict whether a suggested translation is viable. With multiple bilingual dictionaries, a directed graph is constructed where nodes are unique words associated with their language and part-of-speech tag. Depth-first search is applied to find cycles in the graph. Translations found in cycles with a translation in the dictionary from the target word back to the source are considered to be positive examples, whereas translations found in paths but not cycles are treated as negative instances. Additional features are passed on to the model as well, such as the frequency of source word in a dictionary, number of available paths between the source and target words, and, in the case of sharing the language family, the average Levenshtein distance between all the words in the path. This method has been neither investigated nor evaluated for endangered languages.

### Existing dictionaries

Existing dictionaries are written in XML formats. Erzya and Moksha dictionaries follow the XML format of Giella Language Technologies and both are hosted publicly on Github.[2] Figure 1 shows how a Skolt Saami lexeme, *piânnai,* is represented. All translations of the lexeme are listed under the meaning group element, <mg>. For instance, the English and Finnish translations of the word are *dog* and *koira,* respectively.

[2] Erzya: https://github.com/giellalt/lang-sms and Moksha: https://github.com/giellalt/lang-mdf

These dictionaries are available in Akusanat infrastructure,[3] and have been primarily curated by Jack Rueter for over ten years. The dictionaries are multilingual and enhanced with different graph relations to other information.



```
▾<e meta="04" id="gt_N_243">
   <map stamp="gt_N_243:fm_1482:sml_18948" fin_lemma_id="8404::"/>
   <rev-sort_key>iannâip</rev-sort_key>
   ▾<lg>
.....<l pos="N">piânnai</l>
.....▸<inc-sampling/>
.....▾<stg>
.......<st Contlex="N_PIYNNAI" inflexId="1.4">piâ%{'Ø%}n'n</st>
.....</stg>
.....▸<inc-audio/>
.....▸<etymology/>
   </lg>
.....▸<sources/>
   ▾<mg relId="0">
.....▾<semantics>
.......<sem class="ZOO">DOMESTIC</sem>
.......<sem class="HUMAN">FAMILY</sem>
.....</semantics>
.....▾<tg xml:lang="deu">
.......<t pos="N">Hund</t>
.....</tg>
.....▾<tg xml:lang="eng">
.......<t pos="N">dog</t>
.....</tg>
.....▾<tg xml:lang="hun">
.......<t pos="N">kutya</t>
.....</tg>
.....▾<tg xml:lang="fin">
.......<t pos="N">koira</t>
.....</tg>
...
```

*Figure 1.*

*A snapshot of the lexeme element* (piânnai) *in the Skolt Saami XML*

We have imported all of the existing XML dictionaries into Ve′rdd. In total, 118,954 and 55842 lexemes were imported for Erzya and Moksha, respectively, along with their corresponding translations. These dictionaries did

[3] https://akusanat.com

not contain any translations between the two languages, whether it is Erzya to Moksha or the other way around. However, two small dictionaries are available with such translations,[4] which only contain around four thousand translations between Erzya and Moksha.

At the moment, Ve′rdd contains 42 languages, resulting in nearly 2 million lexemes and 1 million translation pairs. In the following section, we describe how we extended the dictionaries' translation candidates automatically based on existing translations for other languages. When conducting an automatic task like this the attention to quality is extremely important, so we present below a thorough evaluation and overview of results.

## Dictionary augmentation

Since our data is in a graph structure in our Ve′rdd system, we can apply graph based methods for the dictionary augmentation. We build our method on an existing link prediction algorithm, Tanimoto similarity coefficient (Tanimoto 1968), which we repurpose and extend for our purposes. This means that Tanimoto coefficient is used to compute a score based on the common neighbours between the source and target nodes with respect to the total number of their neighbours. The mathematical formula for Tanimoto coefficient can be seen in Figure 2, where A denotes set of source language translations and B a set of target language translations.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

*Figure 2.*
*Tanimoto coefficient*

In practice, our method assigns a higher score for new translation pairs that have high support from within the graph structure. Hence, the more translations to other languages a word in Erzya and Moksha have in common, the more likely it is that they are also translations of each other. This way, the output of our system can also indicate how certain the model is that two words are good translations for each other.

We iterate over all potential translation pairs between Erzya-Moksha and Moksha-Erzya, where at least a single translation is shared between the pair. Thereafter, we assess the confidence score using the Tanimoto coefficient

[4] https://gtsvn.uit.no/langtech/trunk/words/dicts/myvmdf and https://gtsvn.uit.no/langtech/trunk/words/dicts/mdfmyv

metric. This process resulted in 43,484 and 43,636 translation pairs for Erzya-Moksha and Moksha-Erzya, in the same given order.

### Evaluation and results

For our evaluation, samples of 300 new translations for both Erzya-Moksha and Moksha-Erzya were evaluated by a knowledgeable linguist. His task was to indicate how good the new predicted translations were. The translation pairs were graded on the basis of three options, which are a perfect match, a near miss and a completely wrong translation. For quantitative analysis, we assign points  from 1 to 3 to the grades, respectively.

The results can be seen in Table 1. We report the values for all 300 samples and a subset of them without multiword expressions as they were problematic. The results indicate that the method excels at finding translations between single words (with 65% and 77% accuracy for accurate predictions); nonetheless, the performance of the method is still acceptable even when including multi-word expressions. When observing the difference percentage of perfect predictions between Erzya-Moksha (37%) and Moksha-Erzya (68%), we notice a difference in the efficiency of the method which is probably due to the fact that the existing dictionaries are independent and of different quality.

| | 1 (perfect) | 2 (near miss) | 3 (wrong) |
|---|---|---|---|
| Erzya-Moksha | 37% | 33% | 30% |
| Erzya-Moksha (without MWE) | 65% | 21% | 14% |
| Moksha-Erzya | 68% | 19% | 13% |
| Moksha-Erzya (without MWE) | 77% | 18% | 5% |

*Table 1.*
*Percentages of the scores assigned to the new predictions*

So-called perfect matches may also include fuzzy equivalent pairs, but there may also be instances where less than desirable but equivalent translations are given. In 1.1–1.3, the matches are perfect, but they may require additional context constraints. In 1.2, we can observe the pair ‘unfound’ vs ‘undiscovered’, where ‘undiscovered’ might denote a more limited interpretation. For example, the English translation for ‘riddle’ in 1.4 shows that Erzya attests to a Russian loanword *загадка* even though a native word *содамоёвкс* is available in Ve′rdd.

1.1 Erzya *чивалгомаёнов* 'westwards' and Moksha *шимадомань шири* 'westwards';
1.2 Erzya *апак муе* 'unfound' and the Moksha *апак панжек* 'undiscovered (lit. unopened)';
1.3 Erzya *эвропань ломань* 'European person (lit. person of Europe)' and Moksha *европэряй* 'European (lit. Europe dweller)';
1.4 Erzya *загадка* 'riddle' and Moksha *содамоёфкс* 'riddle'.

Illustrative examples of the near miss predictions are found in pairs 2.1–2.3. In 2.1 the difference is seen in the body of water indicated. Example 2.2, however, shows a parallel drawn between 'domestic animals' and simple 'animals'. Example 2.3 shows two different parallels drawn, one is between 'daughter' and 'child', whereas the other is between 'foster' and 'not one's own'. This distinction of near miss predictions from fuzzy perfect scores is hard to delimit.

2.1 Erzya *иневедь чире* 'sea coast' and Moksha *ведьжире* 'water's edge';
2.2 Erzya *кудоракшат* 'domestic animals' and Moksha *ракшат* 'animals';
2.3 Erzya *трянь тейтерь* 'foster daughter' and Moksha *аф эсь шаба* 'not one's own child'.

The wrong translations included interesting predictions, where the accuracy of the X-to-English translation may be deemed questionable. In 3.1, for instance, the Erzya *ломанькс путыця* 'venerating, honoring' is being equated to the Moksha 'playing up to someone', hence there appears to be no actual veneration in the aligned Moksha word. Example 3.2, it will be noted, is drawing a parallel between 'international' and 'interpersonal', which would indicate only the 'inter-' segment is shared by the two words. Example 3.3 draws a parallel between 'to show' and 'to confirm', which, although they might be close in some context, are not true matches. Once again, we are reminded of the difficulty of establishing clear parameters for grading automatically predicted translations.

3.1 Erzya *ломанькс путыця* 'venerating, honoring' vs Moksha *мялень ваны* 'respectful (actually: playing up to)';
3.2 Erzya *масторъютконь* 'international' vs Moksha *ломаньётконь* 'interpersonal';
3.3 Erzya *невтемс* 'to show' vs Moksha *кемокснемс* 'to confirm'.

As can be seen from the quality of even the near miss and rejected translation pair predictions, this algorithm may well provide at least a good rating for translation prompting.

## Future work

Ve′rdd has relations other than translation relations to deal with. These include: „Etymological relation” for marking associations between cognates; „Compound relation” for associating compound words with the elements; „Derivation relation” for associating derived words with their source stems and derivational affixes. These could, in the future, be predicted automatically by incorporating methods such as the ones by Alnajjar et al. (2017), Hämäläinen – Rueter (2019) and Alnajjar (2021), test materials could be found in Keresztes (2013).

Etymological relations between words of the Mordvin idioms may be readily augmented in Ve′rdd at the present by entering documentation metadata from sources such as (Keresztes 2011: 107–118), and even more extensive alignments can be found in the Mordwinisches Wörterbuch (1990, 1992, 1994, 1996), both at the lemma level and the etymology level (cf. Rueter 2020).

Future work will involve the enhancement of parallel corpora development for Erzya and Moksha, both literary and vernacular in collaboration with other centers of Mordvin studies. The lexical resources we have created and described here could be used in future in research on morphological derivations, dialect areas and lexical distributions. Furthermore, this approach can be useful for other endangered languages as well.

## Literature

Abramov, K. G. [Абрамов, К. Г.] 2002: Валонь ёвтнема валкс. Мордовской книжной издательствась, Саранск.

Ahlqvist, August 1861: Versuch einer Mokscha-Mordwinischen Grammatik. Forschungen auf dem Gebiete der Ural-Altaischen Sprachen. Commissionäre der Kaiserlichen Akademie der Wissenschaften, St. Petersburg.

Alnajjar, Khalid – Hämäläinen, Mika – Chen, Hanyang – Toivonen, Hannu 2017: Expanding and weighting stereotypical properties of human characters for linguistic creativity. In: Proceedings of the 8th international conference on computational creativity (ICCC'17). 25–32.

Alnajjar, Khalid – Hämäläinen, Mika – Rueter, Jack – Partanen, Niko 2020: Ve′rdd. Narrowing the gap between paper dictionaries, low-resource NLP and community involvement. In: Proceedings of the 28th international conference on computational linguistics: system demonstrations. International committee on computational linguistics. 1–6.

Alnajjar, Khalid 2021: When word embeddings become endangered. In: Multilingual Facilitation. Rootroo Ltd. 275–288.

Donandt, Kathrin – Chiarcos, Christian – Ionov, Maxim 2017: Using machine learning for translation inference across dictionaries. In: TIAD-2017 shared task – translation inference across dictionaries [https://tiad2017.wordpress.com/]. http://ceur-ws.org/Vol-1899/TIAD17_paper_2.pdf

Feoktistov, A. P. [Феоктистов, А. П.] 1976: Очерки по истории формирования мордовских письменно-литературных языков (ранний период). Академия Наук СССР Институт Языкознания, Москва.

Gesese, Genet Asefa – Alam, Mehwish – Sack, Harald 2020: Semantic entity enrichment by leveraging multilingual descriptions for link prediction. In: ArXiv preprint arXiv: 2004.10640.

Hämäläinen, Mika – Rueter, Jack 2018: Advances in synchronized XML-MediaWiki dictionary development in the context of endangered Uralic languages. In: Proceedings of the XVIII EURALEX international congress: lexicography in global contexts. Ljubljana University Press, Ljubljana. 967–978.

Hämäläinen, Mika – Rueter, Jack 2019: Finding sami cognates with a character-based NMT approach. In: Proceedings of the 3rd workshop on computational methods in the study of endangered languages. 1: 39–45.

Hämäläinen, Mika 2021: Endangered languages are not low-resourced! In: Multilingual Facilitation. Rootroo Ltd. 1–11.

Karmanova, A. N. – Jakubiv, T. V. [Карманова, А. Н. – Якубив, Т. В.] 2016: Вежöртас восьтан кывкуд: 3000 сайö кыв. О. И. Уляшев редакция улын. ООО «Анбур», Сыктывкар.

Keresztes, László 1990: Chrestomathia Morduinica. Tankönyvkiadó, Budapest.

Keresztes, László 1999: Development of mordvin definite conjugation. MSFOu 233. Suomalais-Ugrilainen Seura, Helsinki.

Keresztes, László 2011: Bevezetés a mordvin nyelvészetbe. Debrecen University Press, Debrecen.

Keresztes, László 2013: Morféma-alternációk a moksa-mordvin határozott és birtokos személyragozásban. Folia Uralica Debreceniensia 20: 109–152.

Lam, Khang – Al Tarouti, Feras – Kalita, Jugal 2015: Automatically creating a large number of new bilingual dictionaries. In: AAAI conference on artificial intelligence. 2174–2180.

Lam, Khang Nhut – Kalita, Jugal 2013: Creating reverse bilingual dictionaries. In: Proceedings of the 2013 conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies. 524–528.

Maticsák, Sándor 2014: A mordvin lexikográfia évszázadai. In: Maticsák, Sándor – Tóth, Anikó Nikolett – Laihonen, Petteri: Rokon nyelveink szótárai. Fejezetek a finnugor lexikográfia történetéből. Tinta Könyvkiadó, Budapest. 103–130.

Maticsák, Sándor 2017: The beginnings of Mordvin literacy. Buske Verlag, Hamburg.

Mikola, Tibor 1975: N. Witsens Berichte über die uralischen Völker. Aus dem Niederländischen ins Deutsche übersetzt von Tibor Mikola. (Mit einem Anhang.) Studia Uralo-Altaica VII. Szeged.

Pasini, Tommaso, – Navigli, Roberto 2017: Train-O-Matic: Large-scale supervised word sense disam-biguation in multiple languages without manual training data. In: Proceedings of the 2017 conference on empirical methods in natural language processing. 78–88.

Poljakov, O. E. [Поляков, О. Е.] 2004: Мокшень и эрзянь кяльхнень фкакс- и аф фкаксшисна, синь валлувкссна. Эрзянь ды мокшонь кельтнень вейкекс- ды аволь вейкексчист, сынст валлувост. О. Е. Поляков, Jack Rueter. Н. П. Огарёвонь лемса Мордовскяй государственнай университетсь. Саранск: «Красный Октябрь» типографиясь, Саранск.

Rueter, Jack 2016: Towards a systematic characterization of dialect variation in the Erzya-speaking world: Isoglosses and their reflexes attested in and around the Dubyonki Raion. In: Shagal, Ksenia – Arjava, Heini (eds), Mordvin languages in the field. Uralica Helsingiensia 10. Helsinki. 109–148.

Rueter, J. M. 2020: Linguistic distance between Erzya and Moksha. Dependent morphology. In: Клементьева, Е. Ф. – Мочалова, Т. И. – Рябов, И. Н. (ред.), Финно-угорские языки в современном мире: функционирование и перспективы развития: материалы Всероссийской научно-практической конференции, посвященной 95-летию заслуженного деятеля науки РФ, доктора филологических наук, профессора Цыганкина Дмитрия Васильевича. МГУ им. Н. П. Огарёва, Саранск. 90–110.

Russko-mokšansko-èrzjanskij slovar' / Русско-мокшанско-эрзянский словарь 2011. Составители: В. И. Щанкина, А. М. Кочеваткин, С. А. Мишанина. Науч. Ред. Ю. А. Мишанин. Поволжский центр культур финно-угорских народов, Саранск.

Soderland, Stephen – Etzioni, Oren – Weld, Daniel – Skinner, Michael – Bilmes, Jeff 2009: Compiling a Massive, Multilingual Dictionary via Probabilistic Inference. In: Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP. 262–270.

Soosaar, Sven-Erik – Erina, Olga (eds) 2019: Estonian-Erzya Dictionary. [Online.] https://doi.org/10.15155/3-00-0000-0000-0000-07BD8L

Tanimoto, T. T. 1968: An elementary mathematical theory of classification and prediction. Internal IBM Technical Report.

Wiedemann, F. J. 1865: Grammatik der Ersamordwinischen Sprache, nebst einem kleinen Mordwinisch–Deutschem und Deutsch–Mordwinischen Wörterbuch von F. J. Wiedemann, Mitgliede der Academie. Gelesen am 22. December 1864. Mémoires de l'Académie Impériale des Sciences de St.-Pétersbourg, VII$^{E}$ série. Tome IX, № 5. Commissionäre der Kaiserlichen Akademie der Wissenschaften, St. Petersburg.

Wushouer, Mairidan – Lin, Donghui – Ishida, Toru – Hirayama, Katsutoshi 2015: A constraint approach to pivot-based bilingual dictionary induction. In: ACM transactions on Asian and low-resource language information processing (TALLIP). 1–26.

*

## Az erza-moksa szótár automatikus fejlesztése a Link Prediction program segítségével

A veszélyeztetett erza és moksa nyelv kétnyelvű szótárának megalkotására mutatunk be egy újabb lehetőséget, más nyelvű szótárak tanulságai alapján. Ebben a cikkben egy automatikus, adatvezérelt módszert mutatunk be az erza és moksa szókészlet közötti új lexikográfiai kapcsolatok leírására. Arra törekszünk, hogy leírjuk a mordvin kutatások azon lépéseit, amelyek az erza és moksa lexikon szemantikai összehangolását foglalják magukban. Tanulmányunkban röviden bemutatjuk a mordvin nyelvek első szóanyagait és szótárait, amelyek a 17. században keletkeztek. Ennek a lexikonnak a szemantikai fejlesztése fontos a közeli rokonságban lévő, de különálló irodalmi nyelvvel rendelkező erza és moksa közötti nyelvi távolság mérésében. Ez a munka egy nagyobb rendszerbe illeszkedik, beleértve a két nyelv etimológiai, morfológiai és szintaktikai összevetését is.

*Kulcsszavak: számítógépes nyelvészet, erza-mordvin, szótár, szemantika.*

KHALID ALNAJJAR – JACK RUETER – NIKO PARTANEN – MIKA HÄMÄLÄINEN