A free/open-source rule-based machine translation system for Crimean Tatar to Turkish

Memduh Gökırmak
ÚFAL
Charles University
Prague, Czechia
memduhg@gmail.com

Francis Morton Tyers
Department of Linguistics
Indiana University
Bloomington, IN, USA
ftyers@iu.edu

Jonathan North Washington
Linguistics Department
Swarthmore College
Swarthmore, PA, USA
jwashin1@swarthmore.edu

Abstract

In this paper a machine translation system for Crimean Tatar to Turkish is presented. To our knowledge this is the first Machine Translation system made available for public use for Crimean Tatar, and the first such system released as free and open source software. The system was built using Apertium[1], a free and open source machine translation system, and is currently unidirectional from Crimean Tatar to Turkish. We describe our translation system, evaluate it on parallel corpora and compare its performance with a Neural Machine Translation system, trained on the limited amount of corpora available.

1 Introduction

This paper presents a Free/Open-Source prototype shallow-transfer rule-based machine translation system between Crimean Tatar and Turkish. The system is built using Apertium (Forcada et al., 2010), a free and open source platform that facilitates development of rule-based machine translation systems by providing tools that minimize the

The paper will be laid out as follows: Section 2 gives a short review of some previous work in the area of Turkic–Turkic language machine translation; Section 3 introduces Crimean Tatar and Turkish and compares their grammar; Section 4 describes the system and the tools used to construct it; Section 5 gives an evaluation of the system and compares it with a basic neural translation system, also presenting an example of a Crimean Tatar sentence and its translations into Turkish by the systems compared. Finally Section 6 describes our aims for future work and some concluding remarks.

2 Previous work

Within the Apertium project, work on several MT systems between Turkic languages has been started (Turkish–Kyrgyz, Azeri–Turkish, Tatar–Bashkir), but until the release of the pair which this paper presents, the Kazakh–Tatar system (Salimzyanov et al., 2013) was the only one of release level quality, and accordingly the only one released.

Besides these systems and those that are corporately available,[2] a handful of previous works on machine translation systems between Turkic languages exist. MT systems have been reported that translate between Turkish and other Turkic languages, including Turkish–Crimean Tatar (Altıntaş, 2001), Turkish–Azerbaijani (Hamzaoglu, 1993), Turkish–Tatar (Gilmullin, 2008), and Turkish–Turkmen (Tantuğ et al., 2007), though none of these have been released to a public audience. In the development of this system, we use another system developed within the Apertium project, a morphological analyzer for Crimean Tatar (Tyers et al., 2019).

3 Languages

While Turkish and Crimean Tatar belong to different branches of the Turkic family—Oghuz (Southwestern Turkic) and Kypchak (Northwestern Turkic) respectively—historical contact has been intense enough to make the written standards of the two languages somewhat mutually intelligible, although differences in modern vocabulary prevent more complete mutual intelligibility.

[1] http://wiki.apertium.org

[2] e.g., Google Translate, http://translate.google.com

Turkish is the official language in Turkey, and an official language in Cyprus. It is a recognized minority language in Greece, Iraq, Kosovo, Macedonia and Romania. There are around 80 million fluent speakers of Turksih, mostly living in Turkey (Eberhard et al., 2019). Crimean Tatar is a recognized minority language in Ukraine and Romania. There are over half a million speakers of Crimean Tatar, who mostly live in the Crimean peninsula, Uzbekistan, Turkey, Romania, and Bulgaria (Eberhard et al., 2019). The map in Figure 1 shows the two languages' situation among other Turkic languages spoken around the Black Sea.

Figure 1: Location of Turkish (`tur`) and Crimean Tatar (`crh`) within the Black Sea area.

Turkish has undergone a purification process, removing many Arabic and Persian-origin words that it had in common with Crimean Tatar. Turkish has been influenced by and borrowed words mainly from French throughout the 20th century, while the major influence on Crimean Tatar has been Russian. Consider for example the loan word for "bus station", Turkish *otogar* < Fra. "auto- + gare" and Crimean Tatar *avtovokzal* < Rus. "автовокзал".

3.1 Orthographic and Phonological differences

Both the orthographies and the phonologies of the languages are remarkably close, especially in the written standard, but a number of differences are immediately observable at first glance.

The most obvious phonological differences between Crimean Tatar and Turkish are the existence of three phonemes in Crimean Tatar that do not exist in Turkish: /q/, /ʁ/, and /ŋ/.

There are also differences in the treatment of loanwords. Word-final stops at the end of recent loandwords are more consistently devoiced in Turkish, as can be seen in Turkish *mikrop* 'microbe' and Crimean Tatar *mikrob*, or *sülfit* 'sulphide' and *sulfid*. The affricate /ts/ is usually realised as /s/ in Turkish, but preserved in Crimean Tatar. For example, words like Crimean Tatar *tsilindir* 'cylinder', *tsellofan* 'cellophane' tend to appear as *silindir* and *selofan* in Turkish. However, examples such as tsunami do appear in Turkish, and it may also be important that Turkish loans of this sort tend to be of French or English origin, while the Crimean Tatar loanwords are usually from Russian.

3.1.1 Latin script

In recent years, Crimean Tatar has for the most part employed a Latin script almost identical to the Turkish script with the exception of a few letters. The letter *q* is used to represent /q/, a voiceless uvular stop in Crimean Tatar. Neither the sound nor the letter exists in standard Turkish. Crimean Tatar also tends to mark long *a* sounds /a:/ more consistently with a circumflex, *â*, than Turkish, where the character is used sporadically and for more ambiguous purposes — i.e. it can also be used to mark palatalisation. The letter *ñ*, also not used in Turkish, is used for the dorsal nasal /ŋ/, which for the most part no longer exists in Turkish. The use of the letter *ğ* also differs. In Turkish, *ğ* represents what was once a dorsal obstruent, but has since deleted in modern standard Turkish and caused compensatory lengthening of a preceding vowel, e.g. *dağı* [dɑː.ɯ] 'mountain–Poss.3'. In Crimean Tatar, the letter *ğ* represents a uvular fricative /ʁ/, e.g. *dağı* [dɑʁɯ] 'mountain–Poss.3'.

3.1.2 Cyrillic

A Cyrillic alphabet based on that of Russian was used officially from 1938 to the 1990s, and has still not completely fallen out of use today. Unlike some of the other Turkic alphabets, it did not feature special characters that were not present in the Russian alphabet. Consonants and vowels that did not exist in Russian were instead written using digraphs, often involving the hard ъ or soft ь sign.

For example, the consonants represented as *q*, *ğ* and *ñ* in the Latin script are represented as *къ*, *гъ*, and *нъ*, respectively, in the Cyrillic orthography. Also, the vowels represented with *ü* and *ö* in the Latin script are represented with either *уь* and *оь*, or *у* and *о* with a ь after the following consonant, or just *у* and *о* in the presence of certain consonants. See (Tyers et al., 2019) for more details, and how the transliteration module is used to process Cyrillic Crimean Tatar input.

The sentence "Welcome to Crimea!" is shown in

lang. / orthography	text
Crim. Tatar Latin	Qırımğa hoş keldiñiz!
Crim. Tatar Cyrillic	Къырымгъа хош кельдинъиз!
Crim. Tatar IPA	[qɯrɯmʁa xoʃ keldiŋiz]
Turkish	Kırım'a hoş geldiniz!
Turkish IPA	[kɯrɯma hoʃ gʲɛldiniz]

Table 1: "Welcome to Crimea" in Latin and Cyrillic Crimean Tatar orthographies with a Turkish translation.

the Latin and Cyrillic orthographies along with the Turkish translation in Table 1.

3.2 Morphological Differences

The morpheme *-A*, which marks the aorist in Crimean Tatar, serves as the optative mood in Turkish. And while the Turkish aorist *-Ir/-Ar* exists in Crimean Tatar, it is used as a future tense.

Both languages have two basic morphemes for the past tense: Turkish *-mIş* and *-DI* and Crimean Tatar *-GAn*[3] and *-DI*. In Turkish, the distinction is between non-first-hand evidential past and first-hand eyewitness past, whereas in Crimean Tatar the distinction is between non-recent past and recent past. In Crimean Tatar, evidential tenses are usually formed with the additional morpheme *eken*.

Furthermore, Crimean Tatar does not have a distinct strategy for marking progressive aspect, and uses the same morpheme for both non-past and present progressive. In Turkish, the progressive is marked by *-(I)yor*, and can be used with a variety of tenses. Both languages, however, have a progressive construct used in more formal speech and writing, *-mAktA*, which comprises a gerund in locative case.

A number of phonological differences exist between cognate inflectional morphemes in the two lanugages: for example, the Crimean Tatar dative case *-GA*, which can be realised as *-ğa, -ge, -qa, -ke* depending on its phonological environment, corresponds to *-(y)A* in Turkish, realised as *-a, -e, -ya, ye* depending on phonological environment.

3.3 Syntactic differences

Turkish has a richer inventory of morphology relating to relative clauses, particularly verbal adverbs. However, Crimean Tatar exhibits more auxiliary verbs, which are used to add modal and aspectual information to verb phrases.

The languages also differ in their placement of the polar question particle *-mI* relative to person agreement suffixes: in Crimean Tatar the question particle comes after person agreement, whereas in Turkish it tends to come before. For example, in Crimean Tatar *bilesiñ**mi*** 'do you know?' the question particle follows the 2nd person singular agreement suffix *-sIñ*, whereas in the corresponding Turkish form *biliyor **mu**sun*, the question particle preceds the agreement suffix *-sIn*.

[3]The Crimean Tatar morpheme *-GAn* is cognate to the Turkish participle form *-(y)An*.

4 System

The system is based on the Apertium machine translation platform (Forcada et al., 2010).[4] While initially developed to translated between closely related Romance languages such as Catalan and Spanish, the system has evolved to handle different and more distantly related languages. Apertium's code and data are licensed under the Free Software Foundation's General Public Licence[5] (GPL) and all the software and data for the 47 currently released languages (and other pairs being worked on) is available for download from GitHub.[6]

4.1 Architecture of the system

The Apertium translation engine consists of a Unix-style pipeline or assembly line with the following modules (see Figure 2):

- A deformatter which encapsulates format information from the input in superblanks which the other modules process as blanks between words.
- A morphological analyser, implemented as a transducer, which processes surface forms (SF) (words, or, where detected, multiword lexical units or MWLUs) and produces one or more lexical forms (LF) consisting of lemma, part of speech and morphological information.
- A module that disambiguates between possible analyses depending on the context.
- A lexical transfer module which reads each source-language (SL) LF and produces corresponding target-language (TL) LFs by looking them up in a bilingual dictionary encoded as an FST compiled from the corresponding XML file. The lexical transfer module may return more than one TL LF for a single SL LF.

[4]`http://www.apertium.org`
[5]`https://www.gnu.org/licenses/gpl.html`
[6]`https://github.com/apertium`

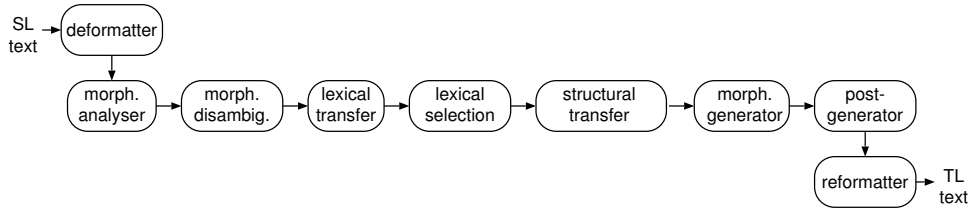

Figure 2: System Architecture

- A lexical selection module which uses rules to choose the best translation of ambiguous source language LFs based on context.
- Transfer rules that work with a shallow method to change grammatical structures in the source language to ones more befitting the target language.
- A morphological generator that produces a TL SF for each TL LF, by applying the correct inflection.
- A post-generator FST to deal with minor orthographic issues.
- A reformatter which de-encapsulates any format information.

The modules are discussed in the following sections.

4.2 Morphological transducers

The morphological transducers are based on the popular Helsinki Finite State Technology (Linden et al., 2011), a free/open-source reimplementation of the Xerox finite-state toolchain. It provides both the lexc formalism for defining lexicons and the twol and xfst formalisms for modelling morphophonological rules. Along with its open-source license, this toolkit is used as it — or the equivalent XFST — has been widely used for other Turkic languages (Cöltekin, 2010; Altıntaş and Çiçekli, 2001; Tantuğ et al., 2006; Washington et al., 2012; Tyers et al., 2012b).

The morphologies of both languages are implemented in lexc, and the morphophonologies of both languages are implemented in twol. Use of lexc allows for straightforward definition of different word classes and subclasses. For example, Crimean Tatar and Turkish have two classes of verbs that have different vowels in the aorist morpheme. Class membership cannot be predicted based on any phonological criteria and is simply a lexical property of any given verb. For example, the Turkish verbs *ısır* and *kır*, "bite" and "break" respectively, inflect differently in the aorist, as *ısır**ır*** and *kır**ar***. Despite the otherwise identical rules of vowel harmony, these two verbs require different paradigms for inflection. This was implemented in lexc with two similar sets of continuation lexica that lead to the appropriate affixes for a given word class.

Twol allows for simple implementation of phonological phenomena such as vowel harmony or voicing/devoicing.

4.3 Bilingual lexicon

The bilingual lexicon currently contains 9,269 stem-to-stem correspondences and was built by:

- Crossing a Crimean-Tatar to Russian + Russian to Turkish dictionary
- Searching for cognates using regular expressions to change frequent differences, e.g. Turkish *hava*, "air", vs. Crimean Tatar *ava*, or similarly *hoca*, "teacher", vs. *oca*
- Consulting a Crimean Tatar to Russian Dictionary manually[7]
- Consulting a Turkish (Ottoman) dictionary[8]
- Adding words provided by Kemal Altıntaş, used in his work on Turkish to Crimean Tatar machine translation (Altıntaş, 2001).

Entries are mostly one-to-one stem correspondences given with their parts of speech, but some also have ambiguous translations.

4.4 Disambiguation rules

We use Constraint Grammar (CG) (Karlsson et al., 1995) for contextual rule-based disambiguation between the possible analyses the analyzer produces for each surface form. The version of the formalism used is vislcg3.[9] The analyzer outputs are fairly ambiguous with an average of around 2.13 analyses per

[7] `http://medeniye.org/lugat`
[8] `http://lugatim.org`
[9] `http://visl.sdu.dk/cg3.html`

crh Sentence	Kerekmey maña öyle feodallar.
Ref. Translation	Lazım değil bana öyle feodaller.
RBMT Output	Gerekmez bana öyle feodallar.
NMT Output	Gerekmez bana böyle otlaklar.

Table 2: Example of MT output for a crh input sentence. Underlined parts of the translation are errors in the output.

form for Crimean Tatar and 2.09 for Turkish. Using the disambiguator, ambiguity is currently down to 1.18 analyses per form for Crimean Tatar and 1.46 for Turkish.

The level of ambiguity has still not converged to near 1, due to many ambiguous affixes that both languages have, particularly in non-finite verbal morphology. However the downside to this is minimized by the fact that the closely related grammar of the two languages means that the very same ambiguity can often carry over in translation without causing an error.

4.5 Lexical selection rules

We use the Apertium lexical selection module (Tyers et al., 2012a).

In some instances, even word translations that are direct cognates may be used in different contexts in the source and target languages. For example, Crimean Tatar *vaqıt* is a word expressing a temporal concept, either a certain point in time or a duration. Turkish has a cognate with very similar meanings, *vakit*, but different contexts elicit different interpretations. Certain collocations such as *bir vaqıt*, "(for) some time", require the use of another translation in Turkish, *süre*. A lexical selection rule to choose the translation *süre* when it occurs with *bir* is written to make sure the correct translation is produced. Similarly the Crimean Tatar word *zümre* has a direct cognate in Turkish, however when it is used in the sense of a language family, it must be translated into Turkish as *aile*, literally "family." The system currently has a total of 13 lexical selection rules.

4.6 Structural transfer rules

Structural transfer rules are written in XML files and are applied left-to-right and longest match first. With equal length matches the preceding rule in the file prevails. There are currently 53 rules for translation from Crimean Tatar to Turkish, and 9 for Turkish to Crimean Tatar.

5 Evaluation

All evaluation was tested against version 0.2.1, or revision 53f133c in the git repository.

5.1 Coverage

Lexical coverage of the system is calculated over freely available corpora of Crimean Tatar. Two years worth of content (2014 and 2015) from Radio Free Europe / Radio Liberty (RFERL)'s Crimean Tatar service,[10] as well as a recent dump of Wikipedia's articles in Crimean Tatar were used.

Corpus	Coverage	Wordcount
Krymr2014	92.6%	874,662
Krymr2015	93.7%	798,666
Wikipedia	89.7%	198,178
Total	**92.8**	**2,032,300**

Table 3: Coverage over corpora. We define coverage here as the percentage of words in the corpus that the system analyzes and produces a translation for.

As shown in Table 3, the naïve coverage of the Crimean Tatar-Turkish MT system over the news corpora approaches that of a broad-coverage MT system, and has less than a tenth of words unknown. The coverage over the Wikipedia corpus is slightly worse, due to the fact that this corpus is "dirtier": it contains orthographical errors, wiki code, repetitions, as well as quite a few proper nouns.

5.2 Translation Quality

Table 2 shows a Crimean Tatar sentence and its translations by both our RBMT system and the NMT system. In both the sentence "I don't need feudal types like that," is translated with *gerekmez* instead of the equivalent *lazım değil*. The RBMT preserves the meaning but doesn't produce the correct vowel harmony in *feodaller*, and the NMT produces the translation "I don't need pastures like this."

We use the metrics BLEU (Papineni et al., 2002) and Word Error Rate, a metric based on Levenshtein distance (Levenshtein, 1966) to evaluate our system on parallel corpora and compare it with the performance of a Neural Machine Translation system trained on the same corpora. We use an *NMT-Small* model from the OpenNMT (Klein et al., 2017) framework for the neural translation. The model we train is word-level, using Byte-pair Encoding (Sennrich et al., 2015).

[10] https://ktat.krymr.com/

To evaluate our system the need arises for parallel corpora. While aligned sentences ready for MT training are not available, a number of academic works published in Turkey provide Crimean Tatar–language text along with Turkish translations. These works are mostly collections of folk tales (Bakırcı, 2010) and selections from Crimean Tatar literature in the Soviet period, from sources including the literary journal *Yıldız* (Atıcı, 2008; Hendem, 2008) and the works of Ayder Osman (Akın, 2014). Other sources deal with the literature of a certain period (Hakyemez, 2007) or social/political phenomenon (Türkaslan, 2015). We align and tokenize the sentences in these parallel corpora using hunalign (Varga et al., 2007) and the tokenizer script provided with the Moses statistical translation toolkit (Koehn et al., 2007). We are in negotiation with the rights holders to release the gathered corpus under an open licence.

Corpus	`crh` Tokens	`tur` Tokens
Yıldız (Volume I)	192,671	190,769
Yıldız (Volume II)	161,047	160,420
Ayder Osman	22,190	21,950
Poverty Literature	23,701	24,185
Folk Tales	84,499	78,998

Table 4: Parallel Corpora. We join together all of these corpora except for the folk tales, and split this in a 90-5-5 split. We use the 5% test portion and the folk tales to test and compare the NMT system to the other systems.

We use all of the parallel corpora listed in Table 4 except for the folk tales in NMT training, randomizing the order of their sentences and splitting them into train, testing and development sets of roughly 90%, 5% and 5% in proportion. This amounts to about 360 thousand tokens for each language in the training corpora, 20 thousand each for the development corpus and again 20 thousand each for testing. The folk tales corpus has a slightly different orthographic system from standard Crimean Tatar, and is non-trivial to convert into the standard. We use this corpus as another test corpus, to compare the performance of our RBMT (Rule-based Machine Translation) and NMT (Neural Machine Translation) systems in situations showing orthographic or dialectal variety.

Table 5 compares the performance of RBMT and NMT on the system, and provides scores for when translation is not done at all in the rows where the System column is filled with "None." The Rule-based system performs better than the Neural system, in both the WER and BLEU metrics. A number of reasons could factor into this. The orthographic and dialectal variety of the texts used in the aligned corpora may have hindered learning and generalization in the NMT system. The RBMT system is to some degree robust to this, as adding frequent variants of frequent words is a simple issue, and one that we frequently addressed while developing the RBMT system on the Wikipedia and news corpora. It should be noted that none of the parallel corpora used for evaluation were used while developing the RBMT system, including the train and development sets.

Corpus	System	BLEU	WER
Test Corpus	RBMT	20.50	54.83%
Test Corpus	NMT	7.88	76.25%
Test Corpus	None	8.29	69.49%
Folk Tales	RBMT	22.07	52.63%
Folk Tales	NMT	2.27	85.11%
Folk Tales	None	9.04	67.87%

Table 5: Evaluation of Translation Quality. "None" simply measures the BLEU and WER scores on corresponding untranslated parallel sentences in each language.

The majority of RBMT errors are mostly due either to mistakes and gaps in the morphophonology components and disambiguation errors or input words being out of the vocabulary. The NMT errors, however, seem to stem from simple lack of data. The figures achieved given only 360 thousand tokens of training data on each side seems to be consistent with experiments conducted in the literature concerning the relation of NMT performance and the amount of data (Koehn and Knowles, 2017). Taken along with the relative lack of standardization of the language, this should account to some degree for the poor performance.

The sheer similarity (and not inconsiderable mutual intelligibility) of the two languages also benefits the RBMT and the scenario where no system at all is used, in comparison to an NMT system that does not have adequate data to encode and decode input text properly.

6 Conclusion

To our knowledge we have presented the first ever publicly available MT system between Crimean Tatar and Turkish, which is available online for use

on Apertium's website.[11] It has near production-level coverage, but is rather prototype-level in terms of the number of rules. Although the impact of this relatively low number of rules on the quality of translation is extensive, the outlook is promising and the current results suggest that a high-quality translation between morphologically-rich agglutinative languages is possible.

We have evaluated our system on an amount of parallel corpora gathered by linguistics departments in Turkey, and compared the performance with that of an NMT system trained on these corpora. The results indicate that even in 2019, it is feasible to use RBMT between closely related, morphologically rich languages when there are not enough resources to train the cutting edge in neural machine translation.

We plan to continue development on the pair; the coverage of the system is already quite high, although we intend to increase it to 95% on the larger monolingual corpora we have — we estimate that this will mean adding around 5,000 new stems and take 1–2 months. The remaining work will be improving the quality of translation by adding more rules, starting with the CG module. The long-term plan is to integrate the data created with other open-source data for Turkic languages in order to make transfer systems between all the Turkic language pairs. Related work is currently ongoing with Kazakh–Turkish, Uyghur–Turkish, Sakha–Kazakh and (Kazan) Tatar–Turkish. The system is available as free/open-source software under the GNU GPL, and the whole system may be downloaded from GitHub.

Acknowledgements

We would like to thank the anonymous reviewers for their insightful comments on the paper and our colleagues Remziye Berberova, Darya Kavitskaya, and Nick Howell, who contributed to the development of the components of the system specific to Crimean Tatar. The work was done with financial support from Google's Summer of Code program,[12] for which we are also grateful.

References

Akın, Serkan. 2014. Review of the stories named "Bizim Gemimiz", "Yıllar ve Dostlar", "Biz Bir Dünyada Yaşaymız" and "Demircinin Teklifi" by Ayder Osman. Master's thesis, Gazi University, Ankara.

Altıntaş, Kemal. 2001. Turkish to Crimean Tatar machine translation system. Master's thesis, Bilkent University.

Altıntaş, Kemal and Ilyas Çiçekli. 2001. A morphological analyser for Crimean Tatar. In *Proceedings of the 10th Turkish Symposium on Artificial Intelligence and Neural Networks (TAINN'2001)*, pages 180–189.

Atıcı, Abdülkadir. 2008. Crimean Tatar Compilations from the Journal Yıldız (Volume I). Master's thesis, Ege University, İzmir.

Bakırcı, Nedim. 2010. *Crimean Tatar Folk Tales*. Kömen Publishing, Konya, Turkey.

Cöltekin, Çağrı. 2010. A freely available morphological analyzer for Turkish. In *LREC*, volume 2, pages 19–28.

Eberhard, David M., Gary F. Simons, and Charles D. Fennig, editors. 2019. *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, twenty-second edition. Online version: `http://www.ethnologue.com`.

Forcada, Mikel L., Mireia Ginestí Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Gema Ramírez-Sánchez, Felipe Sánchez-Martínez, and Francis M. Tyers. 2010. Apertium: a free/open-source platform for rule-based machine translation platform. *Machine Translation*.

Gilmullin, R. A. 2008. The Tatar-Turkish machine translation based on the two-level morphological analyzer. *Interactive Systems and Technologies: The Problems of Human-Computer Interaction*, pages 179–186.

Hakyemez, Betül. 2007. Selected Stories from Crimean Tatar Literature in the period 1928-1937. Master's thesis, Marmara University, Istanbul.

Hamzaoglu, Ilker. 1993. Machine translation from Turkish to other Turkic languages and an implementation for the Azeri language. Master's thesis, Bogazici University, Istanbul.

Hendem, Elif. 2008. Crimean Tatar Compilations from the Journal Yıldız (Volume II). Master's thesis, Ege University, İzmir.

Karlsson, Fred, Atro Voutilainen, Juha Heikkilae, and Arto Anttila. 1995. *Constraint Grammar: a language-independent system for parsing unrestricted text*, volume 4. Walter de Gruyter.

Klein, Guillaume, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

Koehn, Philipp and Rebecca Knowles. 2017. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

[11] `https://www.apertium.org/?dir=crh-tur`
[12] `https://summerofcode.withgoogle.com`

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710.

Linden, Krister, Miikka Silfverberg, Erik Axelson, Sam Hardwick, and Tommi Pirinen, 2011. *HFST–Framework for Compiling and Applying Morphologies*, volume 100 of *Communications in Computer and Information Science*, pages 67–85.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Salimzyanov, Ilnar, J Washington, and F Tyers. 2013. A free/open-source Kazakh-Tatar machine translation system. *Machine Translation Summit XIV*.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Tantuğ, A Cüneyd, Eşref Adalı, and Kemal Oflazer. 2006. Computer analysis of the Turkmen language morphology. In *Advances in Natural Language Processing*, pages 186–193. Springer.

Tantuğ, A Cüneyd, Eşref Adalı, and Kemal Oflazer. 2007. A MT system from Turkmen to Turkish employing finite state and statistical methods.

Tyers, Francis M., Felipe Sánchez-Martínez, and Mikel L Forcada. 2012a. Flexible finite-state lexical selection for rule-based machine translation. In *Proceedings of the 16th EAMT Conference*. European Association for Machine Translation.

Tyers, Francis M., Jonathan North Washington, Ilnar Salimzyanov, and Rustam Batalov. 2012b. A prototype machine translation system for Tatar and Bashkir based on free/open-source components. In *First Workshop on Language Resources and Technologies for Turkic Languages*, page 11.

Tyers, Francis M., Jonathan North Washington, Darya Kavitskaya, Memduh Gökırmak, Nick Howell, and Remizye Berberova. 2019. A biscriptual morphological transducer for Crimean Tatar. In *Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages*.

Türkaslan, Nesibe. 2015. Poverty Literature in Crimean Tatars [sic]. Master's thesis, Gazi University, Ankara.

Varga, Dániel, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2007. Parallel corpora for medium density languages. *Amsterdam Studies In The Theory And History Of Linguistic Science Series 4*, 292:247.

Washington, Jonathan, Mirlan Ipasov, and Francis Tyers. 2012. A finite-state morphological transducer for Kyrgyz. In Calzolari, Nicoletta (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).