Finite-state morphological analysis for Gagauz

Sevilay Bayatli,⋆ Güllü Karanfil,† Memduh Gökırmak,‡ Francis M. Tyers♣
⋆ Faculty of Electrical and Computer Engineering, School of Science and Engineering, Altınbaş University
† Anadolu BİL Meslek Yüksekokulu, Uygulamalı Rusça ve Çevirmenlik, İstanbul Aydin University
‡ Department of Computer Engineering, School of Engineering, İstanbul Teknik Üniversitesi
♣ School of Linguistics, Faculty of the Humanities, Higher School of Economics
sevilaybayatli@gmail.com, gullukaranfil@hotmail.com, memduhg@gmail.com, ftyers@hse.ru

Abstract

This paper describes a finite-state approach to morphological analysis and generation of Gagauz, a Turkic language spoken in the Republic of Moldova. Finite-state approaches are commonly used in morphological modelling, but one of the novelties of our approach is that we explicitly handle orthographic errors and variance, in addition to loan words. The resulting model has a reasonable coverage (above 90%) over a range of freely-available corpora.

Keywords: morphology, Gagauz, finite-state transducer

1. Introduction

Gagauz is a Turkic language of the Oghuz group spoken by approximately 140,920 worldwide and 105,000 in the Republic of Moldova where it is co-official with Moldovan[1] in the autonomous region of Gagauzia. Most speakers are bilingual in Russian and education takes place almost entirely in Russian. This leads to a number of issues with attempting to create a wide-coverage morphological model, as text *in the wild* is often less orthographically accurate than in languages which are widely taught. There is a further issue that keyboards may not be easily available, and so the choice might be between a Turkish keyboard which does not include ţ and a Moldovan keyboard which does not include ı, ö, ü.

The paper is laid out as follows, in Section 2. we describe give a brief typological description of Gagauz, in Section 3., in Section 4. we describe the evaluation and results then in Sections 5. and 6. we give some thoughts on further work and conclusions.

2. Gagauz

To our knowledge to date there has been no computational linguistic work on Gagauz. There are two available grammars, one in Russian (Покровская, 1964) and one in Turkish (Özkan, 1996). The former describes the literary language, where the latter takes a comparative approach, comparing Gagauz with Turkish. There have also been a number of works on Gagauz lexicography, for example Каранфил (2009).

As with other Turkic languages, Gagauz is an SOV language. It has no gender, plural marking and seven cases (nominative, genitive, dative, accusative, locative, ablative and instrumental). Vowel harmony is applied in affixes. Inflectional marking is used to indicate possession and also for agreement within subordinate clauses, *işlediņi gördüm* 'I saw that you were working' (lit. 'I saw your working').

[1] Moldovan is one of the official names for Romanian in the Republic of Moldova. The discussion of if Romanian and Moldovan are separate languages is a sociopolitical one and so for the purposes of this article we consider the names to be equivalent.

Figure 1: Location of the Gagauz speaking area (gag) within the Black Sea region, relative to other Oghuz (Turkish – tur, and Azerbaijani – azb and azj) and Kypchak (Urum – uum, Crimean Tatar – crh, Karachay-Balkar – krc, Nogay – nog, Kumyk – kum and Kazakh – kaz) languages.

Subordination is done principally with verbal affixes which may have multiple possible syntactic readings.

There are many derivational processes, some being very productive, such as the *-k{I}* morpheme which creates attributives from locatives (1a) and substantives from genitives (1b).

(1) a. *Komrat-ta-kı sport liţeyi aç-ıl-dı*
Komrat-LOC-ATTR sport lycée-3SG open-PASS-PAST
'The sports lycée in Komrat was opened.'

b. *ana-m-ın-kı-nı gör-dü-m*
mother-1SG-GEN-SUBST-ACC see-PAST-1SG
'I saw my mother's (one)'

Gagauz shows the usual variation regarding voicing/devoicing of initial consonants in the Oghuz branch of the Turkic family, for example *taa/daha* 'more'. Gagauz also does not exhibit the Turkish characteristic of final consonant devoicing, e.g. the word arab becomes arap in Turkish but does not undergo this change in Gagauz.

Gagauz makes a distinction between the sounds *e* [e] and *ä* [æ], like Azerbaijani and unlike Turkish. Turkish tends to avoid long vowels, specifically in native morphology. Where there would be a voiced velar approximant or a glide in Turkish, in Gagauz the consonant is often not present and there is instead a long vowel. This is not limited to morphology, and some of the lexicon also exhibits this difference, as can be seen in the same example of Turkish *daha* and Gagauz *taa* 'more'.

Like many other languages post-Soviet nations, certain writing conventions have been adopted from Russian, such as a long dash – between the subject and predicate of a declarative sentence with a zero copula. *Arab dili – angisindä laf eder arab halkı* 'The Arabic language – that which the Arab people speak.'

3. Methodology

Development broadly follows the methodology described in Washington et al. (2014), using the Helsinki Finite-State Toolkit, HFST (Lindén, 2009). This toolkit supports the `lexc` formalism for building lexicons and the `twol` for defining phonological constraints. It also supports weighted finite-state transducers. The system is composed of a lexical transducer implemented in the `lexc` formalism, and three phonological/orthographical transducers implemented in the `twol` formalism.

3.1. Lexicon and morphotactics

The lexical transducer, which maps between lexical forms and morphotactic forms (see the example for *insannarına* 'to the people of' in Table 1), was created in the following way: The lexicon was developed completely by hand based on a frequency list. The nominal morphotactics was based on that in Washington et al. (2014), while the verbal morphotactics was created from scratch based on the grammars by Покровская (1964) and Özkan (1996) and corpus investigation. There is a freely-available morphological analyser for Turkish by Çöltekin (2010), but we decided against reusing the verbal morphotactics, as despite Gagauz and Turkish being closely related, there are a number of differences in the verbal morphotactics that would make it difficult to transfer (for example, Gagauz does not have the progressive form that is found in Turkish).

The lexical transducer consists of 5,211 lexemes and 101 continuation classes (sets of suffixes).

3.2. Text encoding

Gagauz uses two letters common with Moldovan, ş /ʃ/ and ţ /ts/. In Moldovan these are normatively spelt as ș and ț using a comma as opposed to a cedilla. In Gagauz text both encodings are found and so we implemented a `twol` file to allow both variants.

3.3. Morphophonology

The morphophonological component is implemented using two-level morphology, `twol` (Koskenniemi, 1984); a total of 24 rules are applied to the lexical forms (see §3.1.) in order to produce surface forms.

For example, in the passive the archiphoneme is -{i}{л}[2], the {л} changes to *n* if it is preceeded by an *-l-*. For example the underlying form of *bulunduk* 'we were found' is `bul-{i}{л}-{D}{I}-k`. The `twol` rule in (2) implements this constraint.

```
(2) "{л} to n in passive following l"
    %{л%}:n <=> :l/:0* %{i%}: _ ;
```

Another rule (3) implements the vowel harmony of the `{I}` archiphoneme:

```
(3) "Vowel harmony for archiphoneme {I}"
    %{I%}:Vy <=> :Vx [:Cns|LowerCns:]+/:0* _ ;
      except
        :Vx [ LowerCns: ] %>: %{s%}: _ %{n%}: ;
      where Vx in ( a â ä e ê o ö u ü ı i i )
            Vy in ( ı ı i i ı u ü u ü ı i i )
      matched ;
```

This rule has an exception for cases of stem consonant elision, for example, the word *topraana* 'to its land' has the underlying form of `topra{k}-{s}{I}{n}-{й}{A}`, where the `{s}` surfaces as null because of the preceeding `{k}`, the `{k}` surfaces as null because it ends up between two vowels, *a* and `{I}` and finally the `{I}` is constrained to assimilate with the previous vowel (carried out by another rule).

3.4. Orthographic errors

Orthographic errors are modelled both in the lexicon and in the phonological rules. In the lexicon errors in stems, for example **iyilik* instead of *iilik* 'goodness', are marked with a comment `Err/Orth`. For typographical, orthographical and phonological errors, such as **içinde* instead of *içindä* 'in the inside of', a separate `twol` file is used for modelling errors. In this error-model `twol` file, we relax the constraints to allow for possible errors, for example `{I}` can surface as {ı, i, u, ü}.

3.5. Compilation

We first compile the lexicon without the orthographic errors to produce the transducer L_n, we then compile the two-level normative rule file to produce the set of rule transducers P, we compose L_n with P to produce the transducer which contains the normative surface forms, T_n. We then compile the full lexicon L_f and compose it with the two-level error rule file E, creating the transducer with all the possible strings T_f (this includes orthographic errors). After that, we subtract the strings in T_f that are in T_n and append a tag to each string indicating orthographic error, `<err_orth>`, and call this the error transducer T_e. The final transducer is the union of T_n and T_e where each pair of erroneous surface form and analysis has a tag indicating it is an error at the end of the analysis.

4. Results

We calculate the *naïve coverage* for the analyser over a number of available corpora: Gagauz Wikipedia, a collection of

[2]We chose Cyrillic {л} 'l' to represent this as Latin 'l' was used for the plural morpheme.

Lexical form		Morphotactic form
`insan<n><pl><px3sp><dat>`	↔	`insan>{L}{A}r>{s}{I}{n}>{й}{A}`

Table 1: Morphotactic representation of the surface form *insannarına* 'to the people of'. The symbols within '{' and '}' characters are archiphonemes which may appear in the surface as a number of different characters, for example `{L}` may appear as *l* or *n*, `{I}` may appear as any high vowel: *ı*, *i*, *u*, or *ü*. The symbols within '<' and '>' characters are grammatical tags, for example `<px3sp>` is the third-person possessive suffix.

Corpus	Genre	Tokens	Coverage (%)	
			GAG	TUR
Ana Sözü	News	525,483	90.2	62.1
Wikipedia	Encyclopaedic	163,403	90.8	64.3
NT	Religion	330,431	90.0	66.2

Table 2: Coverage of the morphological analyser over a range of corpora. The column GAG refers to the coverage using our implementation, while TUR is the coverage after running the TRMorph morphological analyser. As can be seen, despite being closely-related, differences amount to a substantial difference in coverage between the two analysers.

texts from the news site *Ana Sözü*[3] and the New Testament (NT) in Gagauz. Table 2 presents these results. Naïve coverage refers to the percentage of surface forms in a given corpora that receive at least one morphological analysis. Forms counted by this measure may have other analyses which are not delivered by the transducer. We also give the results for running the corpora through TRMorph (Çöltekin, 2010), a freely-available analyser for Turkish, which can be considered a kind of baseline — indicating the performance that could be achieved simply by running a mature analyser for a closely-related language.

In order to evaluate the analyser on a deeper level, We selected 1,000 tokens at random from a list of unique tokens produced from the concatenation of all the corpora. We gave these tokens in 60 characters of context each to a speaker of Gagauz along with the following questionnaire:

- **Is the word Gagauz?** By this we mean could the word be seen in texts in Gagauz, it may be a word with Turkic roots, like *üüredici* 'teacher', or it may be a Russian/international word like *komerţiya* 'commerce' or a word from any number of other languages such as Arabic, e.g. *ceza* 'fine'. The important question is not, "is this a native Gagauz word?" but "could this word be seen in Gagauz texts". A counter example would be *attempt*, an English word which would not appear in Gagauz texts. Another example would be *html* which is not a Gagauz word but some kind of code. Please answer 'NO' if the word is a concatenation of two words caused by a tokenisation error for example *sevincimnänAteş*.

- **Is the word correctly spelt?** By this we mean, is the word spelt correctly according to Gagauz orthography, this includes using the correct diacritics and special characters e.g. *intergraţiya* 'integration' not **intergratsiya* and *lääzım* 'necessary' not **laazim* or **laazım*. Other examples of typographical errors might be adopting Turkish orthographical rules like **Türkiye'nin* instead of Gagauz *Türkiyenin*. Please pay special attention to vowel harmony.

- **What is the part of speech?** Please give the part of speech of the word, choose from: Noun, Adjective, Verb, Adverb, Other.

- **What is the lemma?** Please give the lemma of the word. This is the dictionary form of the word, for example the lemma of *kitaplar* 'books' is *kitap* 'book'.

They then filled out the answers to the questionnaire in a spreadsheet, a sample of which can be found in Table 3.

Out of the tokens, 90 of them were not Gagauz words, and 99 were misspellings. This gave us 910 tokens to evaluate the analyser.

Table 4 shows the coverage of each of the parts of speech according to the test corpus. Note that unlike the naïve coverage, this is a coverage of a random set of tokens and does not take frequency into account. Even so we can see that most of the unknown words come from the open classes (adjectives, nouns and proper nouns).

Figure 2 gives an example of output from the transducer.

5. Future work

As Gagauz is syntactically very close to Turkish, we would like to try cross-lingual methods to morphological disambiguation and dependency parsing. There is an existing treebank of Turkish (Sulubacak et al., 2016) in the format of the Universal Dependencies project and this would be ideal to train a statistical disambiguator. In addition we would also like to explore machine translation between Turkish and Gagauz using the Apertium platform (Forcada et al., 2011). It is worth noting that there are a number of problems in the phonological rules that we are intending to fix. We would also like to expand the lexicon.

6. Concluding remarks

We have presented the first computational model of Gagauz morphology. The transducer has good coverage of a range of available corpora and can handle a range of issues relating to orthography and encoding that are found when dealing with Gagauz. The transducer has the potential to be of great use to the language community as a spellchecker as well as being a key part of other language processing tools. The code is

[3] The web pages was scraped from the home page at `http://www.anasozu.md` in HTML and the text was extracted using a custom Python script

ID	Word	Gagauz?	Spelling?	POS	Lemma	Context
5	payedelecklär	YES	PAYEDILECEKLÄR	FİİL	PAYET-	... dı. Onnar üç uurda proektlara **payedelecklär**. ...
71	ministrulara	NO	–	–	–	... ne premyer-ministruya, ne da **ministrulara** hiç bir soruş ta ...
109	ordenını	YES	ORDENİNİ	İSİM	ORDEN	... likasının "Ordinul de Onoare" **ordenını**. ...
145	süveriz	YES	SÜÜYERİZ	FİİL	SÜÜ-	... nardan çektiimiz için bunnarı **süveriz**. ...
181	İnstitunda	YES	İNSTİTUTUNDA	İSİM	İNSTİTUT	... uzıcescu" adına İncäzanaatlar **İnstitunda** (1972-1992) kulturo ...
…	…	…	…	…	…	…

Table 3: An sample of five lines from the evaluation questionnaire. Note that the Moldovan word *ministru* 'minister' appears here with Gagauz morphology *ministrulara* 'to the ministers'. The annotator has decided that this is a mistake as the normative Gagauz word would be *ministr* 'minister'. Although this would appear to go against our definition of 'is the word Gagauz?' (e.g. appears with Gagauz morphology in a normal Gagauz text) we deferred in all cases to the judgement of the annotator.

```
^Kendi/kendi<det><ref>$
^insannıı/insannık<n><px3sp><nom>$
^için/için<post>$
^hem/hem<cnjcoo>$
^becerikli/becerikli<adj>$
^çalışmakları/çalış<v><tv><ger><pl><px3sp><nom>$
^için/için<post>$
^“/“<lquot>$
^Komrat/Komrat<np><top><nom>$
^kasabasının/kasaba<n><px3sp><gen>$
^şannı/şannı<adj>$
^vatandaşı/vatandaş<n><px3sp><nom>$
^”/”<rquot>$
^adını/ad<n><px3sp><acc><acc>$
^taşıyêr/taşı<v><tv><pres><p3><sg>$
^./.<sent>$
```

Figure 2: Example output of the morphological analyser, the analyses have been disambiguated in context in order to fit on the page. The sentence is *Kendi insannıı için hem becerikli çalışmakları için "Komrat kasabasının şannı vatandaşı" adını taşıyêr.* 'She holds the title "honoured citizen of the town of Komrat" for her skilled work for her people.'

Category	Known	Unknown	Coverage (%)
Punctuation	1	0	100.0
Conjunction	1	0	100.0
Particle	1	0	100.0
Numeral	7	1	87.5
Adverb	10	2	83.3
Verb	164	49	77.0
Pronoun	16	9	64.0
Noun	296	200	59.8
Proper noun	58	68	46.0
Adjective	11	15	42.3
Abbreviation	0	1	0.0
Total:	565	345	61.0

Table 4: Coverage of part of speech categories in the randomly selected test set.

available under a free/open-source licence.[4]

7. Acknowledgements

We would like to thank Alla Büük and Kämil Gagauz for their help as native speakers.

[4] https://svn.code.sf.net/p/apertium/svn/incubator/apertium-gag

References

Çöltekin, Ç. (2010). A Freely Available Morphological Analyzer for Turkish. In: *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC2010)*, pp. 820–827.

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., and Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. In: *Machine Translation* 25.2, pp. 127–144.

Koskenniemi, K. (1984). A General Computational Model for Word-form Recognition and Production. In: *Proceedings of the 10th International Conference on Computational Linguistics and 22Nd Annual Meeting on Association for Computational Linguistics*. ACL '84. Stanford, California: Association for Computational Linguistics, pp. 178–181. DOI: 10.3115/980491.980529. URL: https://doi.org/10.3115/980491.980529.

Lindén, K. (2009). Guessers for Finite-State Transducer Lexicons. In: *Computational Linguistics and Intelligent Text Processing 10th International Conference, CICLing 2009*, pp. 158–169.

Özkan, N. (1996). Gagavuz Türkçesi. Ankara: Ankara.

Sulubacak, U., Gökırmak, M., Tyers, F., Çöltekin, Ç., Nivre, J., and Eryiğit, G. (2016). Universal Dependencies for

Turkish. In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan, pp. 3444–3454.

Washington, J. N., Salimzyanov, I., and Tyers, F. M. (2014). Finite-state morphological transducers for three Kypchak languages. In: *Proceedings of the 9th Conference on Language Resources and Evaluation, LREC2014*.

Каранфил, Г. (2009). Гагаузская лексикология. Комрат.

Покровская, Л. А. (1964). Грамматика гагаузкого язка: Фонетика и морфология. Moscow: Издательство «Наука».