Sorani Kurdish Spell Checker

Rezhwan Kamal ‘23

Computational Linguistics

rkamal1@swarthmore.edu github.com/rezKamal

I used my existing finite-state transducer (developed for the Morphological Analyser lab) to create an open-source spell checker that makes it quicker and easier to type correctly in Sorani Kurdish script using different keyboard layouts.

The Language:

Central/Sorani Kurdish (کوردی) is an Indo-Iranian SOV language with rich functional morphology and agglutination. Many of its ~10 million speakers are fluent in Arabic and/or Farsi. It is worthy of note that there already exists a ckb package on Apertium, which includes a transducer written in .dix format.

These points will become relevant when discussing implementation and design considerations.

Background:

Given the size of the Sorani community, there is no shortage of authentic material on the Internet for the language. The main corpus I worked on this semester was Sorani Kurdish Wikipedia, scraped using the Python Wikipedia Extractor tool. With ~2 million words, it opened the door to working on projects which by nature require evaluation over large corpora. From that subset of projects, I chose to improve my transducer and create a working spell checker.

The motivations behind this choice were twofold:

- As a native Sorani speaker, I can implement morphological rules via lexd and twol without needing to do much research.
- I wish to implement a spell checker that improves the typing experience of those who either (1) misuse the Shift key on the Kurdish layout or (2) use a different layout altogether to type Kurdish. In my experience, most Kurdish typists are in one of these two groups, and a spell checker that corrects for these habits can encourage more people to type in Kurdish.

My principal intent in releasing my package is for it to be used as an additional reference for any linguist looking to build or build upon a transducer or spell checker for Sorani Kurdish.

Methods:

Implementing grammar points in lexd

Example Sentence

توپی کورِه بچووکەکەم بردەوە → I brought the little boy’s ball back.

توپ	-ی	کورِ	-ە	بچووک	-ەکە	-م	برد	-ەوە
ball	-of	boy	-of	small	-DEF	-I	brought	-back/again

Output

```
توپ<n><sg>+ی<pr> ↔ توپی
کور<n><sg>+ە<pr> ↔ کورە
بچووک<adj><def><sg>+من<prn><pers><p1><sg> ↔ بچووکەکەم
بردن<v><tv><past>+ەوە<pr> ↔ بردەوە
```

You can note how agglutinative Kurdish is in this sentence, with every word consisting of 2 or more morphemes. Sometimes, these morphemes are tied to the word itself, as in the case of definiteness markers. In other cases, such as that of an Izafa ending or a subject pronoun suffix on the object, the morpheme(s) added were unrelated to the word itself, and I used the ‘+’ tag to link them together.

Improving the transducer’s coverage

For the Polished RBMT lab, I doubled the coverage of my transducer by:

- **Adding new stems**: I used a script to scrape all the lemmas from Apertium’s ckb-eng transducer.
- **Adding to existing patterns:** I did not need to do much in terms of adding new grammar rules, but given the Sorani’s agglutination, I got a lot out of adding more optional prefixes and endings to existing patterns, such as definite markers on adjectives, Izafa endings on prepositions, etc.

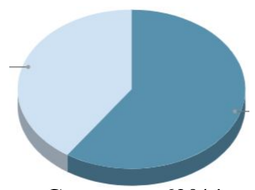
Creation and design of the spell checker

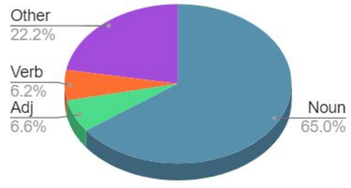
Setting up the spell checker required remote installation of libvoikko. I tested spelling using a shell command which (initially) outputted C for correct and W for incorrect spellings.

To get suggestions (S) for errors, I mapped likely error pairs and added their respective weights. The code snipped shows that the pairings ر/ڕ and و/ۆ are likely to be mistaken for one another. Both mappings are assigned “lighter” weights of 0.5 than the default 1.0. Most of my mappings were between characters corresponding to the same keys on the most common Kurdish layout, or Arabic characters which users sometimes type in place of Kurdish ones.

```
0.5    ڕ    ر
0.5    ۆ    و
```

Evaluation:





Coverage: 63%*

1916 entries

Precision: 98% **Recall: 68%**

*Coverage reported is on the Wikipedia corpus (~2M words). On the Bible corpus (~1000 words) it was ~73%. I believe this difference is down to the frequency of proper nouns and non-Kurdish words in Wikipedia compared to in the Bible.

```
C: توپی
C: کورە
C: بچووکەکەم
W: بردەووە
S: بردەوە
S: بردمووە
S: بردتووە
S: بردووە
S: بردمەوە
```

Conclusion:

The results show that the transducer and spell checker can serve as good scaffolding for another linguist looking to build or improve upon tools for Sorani, with the caveat that there is still much work to be done in terms of improving the transducer’s coverage.

Further work can include:

- Expanding the verb and adjective lexicons to improve coverage
- Including more/narrower twol rules that remove incorrect forms so that the spell checker does not give false negatives.
- Adding paradigms for tenses in the subjunctive mood.
- Somehow getting data on which misspellings are the most common, in order to improve spell checker suggestions.

Acknowledgements:

- Tyers et al., for open-sourcing the Apertium ckb-eng repository.
- Prof. Knerr (CS), for installing libvoikko on the Swat machines.
- Prof. Washington, for supervising my progress in the course.
- Daniel Swanson (course TA), for answering my many questions.
- Trinh Nguyen, for providing the template for this poster.