Encouraging Language Students to Contribute Inflection Data to Wiktionary

Zachary Kurmas
School of Computing
Grand Valley State University
kurmasz@gvsu.edu

ABSTRACT

We propose building a computer program to simplify access to the inflection (i.e., "word ending") data in Wiktionary. This program will make it easier to both (1) look up a word's inflections and, more importantly, (2) edit incorrect inflections. We expect that such a program will encourage foreign language students to both use Wiktionary as a resource and contribute inflection and other grammar data to Wiktionary. We believe that the resulting additional activity will make Wiktionary a better resource for students — especially students of those languages for which there are no cheap, comprehensive inflection resources — and provide data that will be beneficial to the wiki research community.

Categories and Subject Descriptors

H.5.3 [**Information Interfaces and Presentation**]: Group and Organization Interfaces

Keywords

wiki, Wiktionary, language, inflection

1. PROPOSAL AND MOTIVATION

We propose building a computer program to simplify access to the inflection (i.e., "word ending") data in Wiktionary.

Learning a foreign language is challenging – especially for students of languages with complex inflection structures (i.e., languages for which the words change significantly based on their use in a sentence). Most dictionaries contain root words only and provide only minimal (if any) information about how the word should inflect. For example, Polish verbs follow one of seven different conjugation patterns. A typical Polish dictionary will tell a student that the infinitive *mieszkać* conjugates to *mieszkam* in the first person. It then assumes that the student can correctly ascertain which of the seven rules to apply, then correctly apply that rule to generate the desired conjugation. Discerning and applying these rules is more difficult that it appears. Thus, it would be beneficial if a student could simply enter *mieszkać* and explicitly see each of its conjugations.

Polish is one of only a few languages that has an inexpensive, electronically accessible, authoritative source of inflection data. *Słownik gramatyczny jęnzyka polskiego* (SGJP: The Grammatical Dictionary of the Polish Language) is an electronic dictionary containing the inflections of over 245,000 Polish words [2]. In addition, the University of Pittsburgh maintains an online dictionary that also contains inflection tables for many Polish words.[1] The authors of SGJP are aware of similarly comprehensive references for only three other languages.[2]

Wiktionary has the potential to become an SGJP-like grammar reference for many languages (although the data would, obviously, carry less authority). Contributors are increasingly adding inflection data to the English Wiktionary.[3] Of the 7000 Polish words in the English Wiktionary, 4800 contain inflection data. Wiktionary also contains some inflection data for many other languages including Russian, German, Greek, Czech, and Latin. Assuming sufficient accuracy, Wiktionary can potentially become a valuable resource for students of those languages without inexpensive, authoritative sources of inflection data.

It is not difficult to look up inflection data in Wiktionary. Users need only enter
`http://en.wiktionary.org/wiki/WORD#LANGUAGE`
(where *WORD* is the desired word, and *LANGUAGE* is the desired language). However, this interface is not designed specifically with students in mind. A custom program could improve the Wiktionary interface by

- presenting the information more directly,
- maintaining a history of commonly referenced words,
- offering suggestions when words are misspelled,
- automatically correlating words with their root (e.g., suggest looking up "to be" when a user searches for "am" or "are"), and
- providing assistance entering foreign characters on computers without a foreign keyboard configured. (Currently, to look up *mieszkać* in Polish, a user must enter this somewhat awkward url: `http://en.wiktionary.org/wiki/mieszka%C4%87`.)

[1] See `http://polish.slavic.pitt.edu/~swan/beta/`.

[2] http://www.info.univ-tours.fr/~savary/Polonium/Papers/prezentacja-SGJP-Tours.pdf

[3] The English Wiktionary contains more Polish inflection data than the Polish Wiktionary.

WikiSym'10, July 7-10, 2010, Gdańsk, Poland.

The more important benefit of developing a custom interface to language data in Wiktionary is that we can simplify the process of updating the inflection data. Currently, entering new inflection data or correcting existing entries requires the contributor to know the syntax of the MediaWiki template. For example, A user attempting to edit the conjugations of the Polish verb *kupić* would see this:

```
 ====Conjugation====
{{pl-conj-ap|kupić|kupię|kupi|kupią|kupi|kupi
|kupi|kup|ip=kupiono|pp=kupion|pp2=kupieni
|vn=kupienie|ap=kupiwsz}}
```

Editing this entry requires understanding both the concept of MediaWiki templates and the syntax of the `pl-conj-ap` template. We expect that few students — especially those in the process of completing an assignment — will take the time to learn the necessary template syntax to make a correction.

In contrast, students using a custom program could simply enter each correction in a text field and press an "Update" button. The application would then automatically generate the proper template and upload it to the proper place in the Wiktionary article.

2. EXPECTED BENEFITS

We anticipate several benefits from our proposed program:

- It will simplify access to the inflection data in Wiktionary, thereby allowing students to study and complete assignments more efficiently.
- It will simplify the process of correcting mistakes or adding new data, thereby increasing contributor activity. We expect that this increase in activity will increase the number of Wiktionary entries that contain inflection data and improve the overall correctness of the inflection data.

The increased activity should not only improve the utility of the data set for students, but it may also make the data set more interesting/useful to those researchers who study wikis. The increased activity may also make the Wiktionary data more useful to those who study languages and grammar.

3. CONCERNS / RESEARCH QUESTIONS

Encouraging student use of Wiktionary inflection data may have some negative consequences.

- Simplifying the retrieval of individual word forms may discourage some students from learning rules they would otherwise memorize.
- It is not clear whether the data will be sufficiently accurate for use as an educational resource [1].
- It is not clear whether simplifying the process of editing inflection data will result in more correct data, or if it will encourage less knowledgeable or less dedicated users to enter data, thereby lowering the quality of the data.

Each of these concerns above is an interesting research question.

4. DISCUSSION

The purpose of this poster is to facilitate discussion. We want to hear attendees' opinions about the following:

- Which languages would potentially benefit from our proposed program? Some languages don't have a complex enough inflection system to benefit from such a program. Other languages (including Polish) already have an authoritative reference. (We are especially interested to hear about authoritative references similar to [2].)
- Which features will be most important in attracting student use?
- Which of the concerns from Section 3 are valid? Which have already been shown not to be an issue?
- How can we mitigate those concerns that are valid?
- On what platform should we build the program? Should it be a web application? A stand-alone program? An iPhone application?

5. CONCLUSION

We believe that building a custom interface into the inflection data in Wiktionary will (1) help students of foreign languages, (2) increase the amount and accuracy of inflection data in Wiktionary, and (3) possibly provide useful data for those who study wikis and grammar. We look forward to discussing our proposal with WikiSym attendees.

Acknowledgments

We would like to thank the many faculty and students at Grand Valley State University who contributed to this project including: Robert Adams, Christian Trefftz, Jerry Scripps, Christine Rydel, Eva Lewak, Ira Woodring, Ryan Lammers, Sean Fisk, Chris Sechler, and Nick Butcher.

6. REFERENCES

[1] Z. Kurmas. Zawilinski: A library for studying grammar in wiktionary. In *WikiSym '10: Proceedings of the 6th International Symposium on Wikis and Open Collaboration*, New York, NY, USA, 2010. ACM.

[2] Z. Saloni, W. Gruszczyński, M. Woliński, and R. Wołosz. *Słownik gramatyczny jęnzyka polskiego*. Wiedza Powszechna, 2007.