RWC Tool Requirements Review

10 July 2019 / 1:30 PM / **https://zoom.us/j/8333825275**

Attendees

Verna Stutzman, Kevin Warfel, Beth Bryson, Jason Naylor, Sam Delaney, Simeon Eberz, Paul Nelson?

Agenda

Prayer

Project Vision

Reduce the time and effort to get from a rapid word collection workshop to a usable dictionary.

Project Status

Sam Delaney reports on current status

Review Requirements Document

https://docs.google.com/document/d/1GKCmkYh8GmDC2kjqa_YCHtGQjkShvbuIWbLEveBa62k/

Discuss the new features and priorities

Discuss how the tool can be used as part of the process

Discuss the existing features that must be part of the tool, and what can be done in other tools

Discuss possible future requirements

Notes

(If anyone finds any more action items in this writeup, please add them at the end as well. -BB)

STATUS REPORT:

There are two main parts of the tool: data entry, and cleanup.

For data entry, they want to allow both a paper process and paperless.

They envision one screen where a typist could just enter data, and another view that allows separate screens for scribe and glosser to enter their data separately. So far have been focusing on the typist view, because it combines everything, but they plan to separate it out once the functionality is done.

The cleanup part requires a lot more effort from the dev team than the data entry part, so that is where they have focused their time. They have identified 7 "tools" to help with cleanup. They have focused on the two they thought would have the biggest impact: Merge Duplicates, and Create Character inventory. Beth saw both tools, and they have made tweaks to both since Beth last tested.

NEW FEATURES 1

Biggest area of disagreement: need for typeahead for vernacular forms.

Dev concerns:

- Concern about losing data.
- Principle of UI design: If you present a user with an option, they will choose it, even if it is wrong.
- How useful would it be to present them just a vernacular word, without a gloss or SemDom attached to it? Doesn't that increase the likelihood they will choose a word that is in fact different from what they intended to type?
- Someone suggested that "alternate spellings" or "mistypings" could be valuable data for research about the orthography.
- Believe that if the cleanup phase can be reduced from one week (or multiple weeks) down to two days, it won't matter that duplicates were entered.
- Perception that most of the cleanup has to do with other kinds of "garbage", rather than duplicates.
- Understood that the typists were the ones who had some spare time in the workshop, so they could use some of that spare time for cleanup. (Kevin/Verna confirmed: It is the glossing that is the bottleneck, not the typing.)
- Don't want to interrupt the flow of them getting their whole page entered.
- TLC is very concerned with tracking every last piece of data, and not losing any possible valid word.

Responses from DLS:

- Out of 15,000 words collected, around 5,000 are culled during the cleanup phase. Most of this is due to duplicates. (Kevin was having a hard time thinking of other reasons to remove a word during cleanup other than merging it or deleting it because it was redundant.) In general, after cleanup they end up with ⅔ as many distinct senses as the number of raw words that were collected.
- Normally a week is dedicated to the "cleanup phase" in a workshop, often immediately following collection, but sometimes with a break in between. Often the better part of the first day of that goes into finishing the typing. Desire to do more productive things with that time than to merge duplicates that could have been prevented in the first place.
- Even in a paperless workshop, the glossers are working at the same time, and usually a duplicate doesn't show up (for instance, in a different domain) until half a day after it was first entered, so it should have a gloss by then. Seeing the word plus gloss would help the typist know whether it was indeed the same word or not.

- Verna would be disappointed to not get the typeahead feature simply because two scenarios out of 50 don't want it.
- Beth asked if there wouldn't be an option to turn off the vernacular typeahead feature--that way it would not be required for all workshops.
- Kevin noted that although the process of finding duplicates in FLEx is painful, it isn't as painful as they might think, so the possibility for improvement there is not as great as one might think. The mechanics of making the changes is more challenging, but there is still the sense that the greatest payoff for improvement lies in preventing the duplicates from getting into the database in the first place, so they don't need to be cleaned up. The hope is to use the cleanup time on more fruitful endeavors, like glossing or doing better sense differentiation.
- Verna noted that one of the requirements for even doing a RWC workshop is that there already be a stable orthography.
- There are far less expensive ways of getting data about "alternate spellings and mistypings" than conducting a RWC. The sheer volume of data that comes from a RWC makes it difficult to work with, for those purposes.
- Kevin commented that those who have complained about the amount of time needed for the cleanup phase would not perceive this tool as an improvement if it didn't have a way of preventing the duplicates from coming into being in the first place. It would not increase their motivation to do more RWC workshops.

  Conclusion: Jason made an executive decision that they will add a typeahead feature for the vernacular. He agreed that making it optional was a compromise that would address the concerns they had about it.

More new features 1

- Dev team has started dev work on a typeahead for the Glossing portion already. Rather than "typeahead" it is more of an auto-complete, where it will show a drop-down with possible ways to complete it, including Form/Gloss/SemDom of the options. [Beth doesn't understand the difference between "typeahead" vs "autocomplete".]
- They can make the tool guess if the user has put something in the wrong SemDom (e.g., if it was in two domains where one is more specific, they could drop the more general). [Beth has doubts that the facts are quite that simple--wants to discuss with Verna/Kevin more about possible scenarios. Believe giving the user options in the same way they will have options about glosses is better than making it all automatic.]
- There will be a glossing spell checker that would show while they are entering a gloss.
- The tool is built to be localizable.
    - Will need to get it up on CrowdIn so Verna can find people to translate the UI.

    - The UI language and the glossing language are separate (i.e., may not be the same).
    - Multiple glossing languages are allowed.
- The tool will preserve the order that words were entered. [If this appears to be not working, please bring it to Jason's attention, because their belief is that it already is.]

NEW FEATURES 2

- Tagging each piece of data: already in the tool. Tagging for who provided the Form, Gloss, SemDom. What to do with it when it goes to FLEx? Proposal: concatenate the data and put it into the "Source" field in FLEx (a sense-level freeform field).
- Is there a need to record who did which Edits on a record? The tool keeps an entire history of what happened. There isn't a way to represent that in FLEx, so at this point it would be lost when the transfer to FLEx happens, but it would be possible to build an export of that history so it could be used by some other process, or when someone figures out a meaningful way to make use of it.

EXISTING FEATURES TO KEEP

- The "enter by domain" part is nearly done. They are making it so the user can choose whether Form goes into Lexeme Form or Citation Form, and Meaning goes into Gloss or Definition.
- Defaults:
    - Currently always create a new word, not add a sense.
    - Kevin confirmed that FLEx's default (create a new sense, not a homograph) is a better default; that is true for more of the cases.
    - There could be a way to flag cases where a default was chosen, so they could be reviewed later. (Beth was wondering how this flag is any different from, "Find all homographs" or "find all words with more than one sense".)
- There will be a way for the typist to flag words to come back and review them later. (DLS had talked about it, but didn't get it into this requirements doc. This is good.)
- The transfer to FLEx is via a LIFT export.
    - It is pretty straightforward. (Yes, could visualize doing this every day during the workshop.)
    - Question about doing this to the same database each day (mergin), or importing into a fresh empty one each day. If merging into the same one, dev team would need to keep deleted words in the LIFT file and mark them as deleted, so they would in fact be gone when merged with the existing database.
    - Alternative: If the guidelines for the workshop say to use that LIFT file to create a new FLEx project each time, then there would not be a merge issue. Verna noted the team just needs to have a backup of an empty project that they can restore each time and then import into, so they are not creating a new project every day (just restoring).

    - Jason decided to go with this plan: They will not change how they treat deletions in the LIFT file, but count on the participants being instructed not to merge the LIFT file with an existing db.
    - Kevin asked how this would work when there was a db before the workshop, with data in it, and what if the workshop people accidentally deleted something that wasn't supposed to be. [I missed the answer.]
- Publish daily (printouts or Webonary)
    - Exporting to LIFT and creating a new FLEx project will hopefully be easy enough that this can be done from FLEx whenever desired, even on a daily basis or more.
- Browse view of the data: haven't worked on it yet, but expect it to be trivial, given the tools they are working with.
- Dialect flagging: Had not planned on this.
    - In Nepal, a workshop tried to have only one dialect, but discovered the participants came from half a dozen subdialects. Had to clean up that mess after the workshop, and it took months.
    - There may be situations where it is desired.
    - It may be that "dialect" can be linked to a "person", so something could be set at the beginning of a "batch": "this group produced data from dialect A" or "this person produced data from dialect B".
    - It may or may not need to be set on a per-word basis.
- Scientific name comes up because they may be using an encyclopedia or picture dictionary to find the glosses for animals and plants. Such a resource probably also gives the scientific name, so that is an ideal time to go ahead and enter it. (Team agreed to add this field.)
- Etymology: In previous workshops, sometimes the typist will note the Source Language for a borrowed word in parentheses after the gloss:
    - eskirbir 'to write' (Spanish)
    - polisu 'police' (English)
    - Team agreed to think of a way to allow this. (Source Language and Source Form, or only Source Language? It is entry level field in FLEx.)

POSSIBLE ADDITIONAL FEATURES

- Have the custom cleanup tool already
- Audio: planning; not sure if they have started it or not
- Custom SemDom list: had planned for it, but so far just using the FLEx one.
    - Since FLEx doesn't allow custom SemDom lists, to do this a team would need to create a custom list in FLEx and populate it with their custom SemDoms, rather than modifying the official SemDom list in FLEx.
    - If they did that, then when exporting that db to LIFT and importing to this tool, the custom list would come over, with its range sets.
    - It would be minimal effort to allow a custom list like that to be used.
- Collection from a word list

    - Could be implemented by saying the SemDom was “none” and using existing functionality.
    - Would want to make this a non-obvious feature. Need “special knowledge” to know how to turn on that option.
    - Team may look into using the CAWL list. (Beth recommended looking in the WeSay source for that list. Jeff Shrum is familiar with the various people who claim to have the “authoritative copy” of it. Rod Casali, at CANIL, is one.)
    - Verna mentioned there is also an Oxford list.
    - It should also be possible for the user to import their own list.

DREAMS FOR THE FUTURE

- Dev team pointed out that, although this is designed for an “appliance”, it is written as a web app, so if there were a workshop in a place with good connectivity, this could be run over the Internet.
- Detect complex forms: That is a nice dream. Doesn’t seem to fit with a RWC workshop, but it is a good dream to think about.
- Knowledge of paradigms, to help with well-formed citation forms.
    - Had thought of a tool to help with this.
    - Also thought of a tool to help clean up alternations like “eat, to” vs. “eat” vs. “to eat”

THE SEVEN TOOLS

They identified 7 tools that could help with cleanup. That list might be something like:

- Merge duplicates
- Glossing language spell checking
- Build character inventory
- Finding words with characters not in the inventory
- Capitalization (e.g., ask “Is this a proper noun, or should we make it lower case?”)
- Table view (browse view)
- Bulk editing?

Action Items

Implement optional typeahead for vernacular
Use source field to add the user who added the form, gloss, semantic domain, and part of speech
Export entire history
Allow entering scientific name
Design capturing etymology
Design dialect flagging if possible