Requirements for RWC appliance

Beth Bryson
Verna Stutzman
Kevin Warfel
July 8, 2019

A. New Features: Value Add (Tier 1)

These are the “value add” components of the new tool. Without these, there is no point making a new tool, in contrast to enhancing the existing tool.

1. Prevent bad data from getting into the database in the first place.
    a. Rationale: If we are collecting 15,000 words over two weeks, and roughly 5,000 will be thrown out because they are garbage, preventing those from ever being in the database will give us a huge leg up for the cleanup phase.
    b. Caveat: Evaluating the data as it is being typed might slow down the process, but we believe with a sufficiently well-designed tool, it won’t be a big hit. The workaround would be to try to hire more typists. Often it is an outsider (educated national or foreigner) who is doing the typing anyway, so it should be easier to find more typists than those who suggest words. And if the typeahead mechanism is good, then there will be a lot fewer keystrokes for the words that are duplicates.
    c. Two scenarios: Data recorded on paper during collection vs. data typed directly into the computer in the group.
        i. In first scenario, the paper serves as a backup in case of a mistake, so it’s okay for the typist not to type in the duplicates.
        ii. In second scenario, there is opportunity to discuss, so as not to lose info.
        iii. Possibility: The user starts typing a word, then they see the one they want. Preserve as much of the line as they typed, even though they didn’t finish typing the line. Preserve it as “grayed out”? Preserve it only during the typing phase, and let it go away then? But we can’t really think of why this would be wanted. (In a paper workshop, they would draw a red line through it, so that is the record.) Conclusion: We don’t need this.
        iv. The cost of missing a few words is small compared to the pain of having hundreds of obvious duplicates that shouldn’t have gotten into the data in the first place.
    d. Functionality: As the typist is beginning to type a word, they need to see other entries that are similar to it (Form, Gloss, SemDom). If what they are starting to type is already there, they can just choose or edit an existing entry, rather than finishing typing what they were starting to type.
    e. Key feature: Tool has intelligence to propose items that might be duplicates:

    i. Form is misspelled in old or new entry; needs to be corrected.
    ii. Form is different, and represents a new entry.
    iii. Forms are the same, and either: (1) glosses are slightly different (either old or new one is spelled wrong, or (2) there is (correctly) more than one way to gloss it. Might result in a new sense, or concatenated glosses in one sense.) (Should the tool have different behavior if the items are "slightly different" vs "very different"?)
    iv. Glosses in one WS are same, but in other WS are different.
    v. Forms are same, glosses are same (or different glosses are getting merged, as above), but SemDoms are different. Might result in: using both for this sense, choosing old or new SemDom for this sense, or making a new sense with same gloss but different SemDom.
f. Observation: When evaluating "are two items the same?" there are actually three possible answers: (1) identical, (2) similar but slightly different (e.g., "small" vs. "smal", "eat" vs. "to eat" vs. "eat, to", SemDoms that are "more specific" or "less specific" than another), and (3) completely different (e.g., "small" vs. "little", "small" vs. "young", SemDoms in completely different domains). It may be that "slightly different" items result in different possibilities than "completely different" items.
g. Specific things the user needs to be able to do as they are starting to type in an entry, and the tool has proposed some existing entries that may be duplicates:
    i. Choose existing entry (as is; nothing added or changed). What they were about to type is discarded.
    ii. Edit the Form in an existing entry (Leave all else the same; discard what they were going to type. Or this may be in addition to any below operation.)
    iii. Choose existing sense (gloss + SemDom list) (discard what they were going to type)
    iv. Add another gloss to an existing sense. (Add the gloss they are typing, separated by a semicolon, so that separate reversals will be created automatically when we get to that Bulk Edit stage. Add the SemDom, if the current one is not already in that sense.)
    v. Edit the gloss of an existing sense (either instead of what they were going to type, or in addition to adding that)
    vi. Add a new SemDom to an existing sense (discard what they would have typed, but keep the SemDom of the current question)
    vii. Remove a SemDom from an existing sense (in addition to whatever else they are doing)
    viii. Add a new sense (what they are typing becomes the new sense)
    ix. Ability to turn off the interactive mode, and use hard-coded defaults (see "Existing Functionality" section for defaults)
h. New thought: Do we need more discussion about what the "interactive mode" looks like at different points of "typing an entry"? Crucially, is there different

behavior “while they are typing something” vs. “after they finish typing (or accept a proposed completion”? That distinction applies to both the Form and the Gloss.

i. When starting to type the Form, each character results in suggestions of other vernacular words that start with the same character(s). Is this a fuzzy match or exact match? Needs to show glosses along with vernacular; how does that work with the volume that will be possible at the beginning of a word?
ii. After they have finished typing the Form (or chosen something from the autocomplete), this would be a really good time to do the “fuzzy matching” and present as complete a list as possible of all the “entries and their senses” that might have the same Form.
iii. If they didn’t choose anything in that mode, and they start typing the Gloss, what happens?
    1. Autocomplete presents options for just the gloss, for each character they type? Is it based on a spell checker, or existing glosses in the db, or both? Exact match or fuzzy match?
    2. Is there “entry matching” going on, where it tries to do a fuzzy match of both the Form and the Gloss, or would that happen only after they either “finish typing the Gloss” or “choose a gloss from autocomplete”?

2. Glossing language spell checker (this is an additional way to prevent bad analysis language data from getting into the database)
    a. At time of entering
    b. At time of committing words: opportunity to interact with spell checker before committing.
3. Preserve order of words entered until that page is committed
    a. Rationale: If a typist is typing a whole sheet of words, they might lose their place. They want to compare what is on the sheet with what is on the screen to determine which words they have already entered. This is only possible if the screen has the words in the same order as they are on the sheet.
    b. Functionality:
        i. It is okay if the words get alphabetized once the typist moves on to a new domain.
        ii. Ability to switch between various sort orders: Order entered, alphabetical, longest/shortest, from end.

B. New Features: Value Add (Tier 2)

These are important new features, but they are less essential than the Tier 1 features.

1. Real-time multi-user access to the same database

    a. Rationale: If typists and editors are working in the same database (without having to do Send/Receive), they are changing live data, rather than data that may have been adjusted by someone else since they last pulled the data over.
    b. (This isn't a current pain point, but this may enable the other desired features.)
2. Tag each piece of data for the person who provided it
    a. Rationale: The reliability of the data depends on the person who touched it (degree of knowledge of the language, insider/outsider, different dialects)
    b. Functionality: Need to tag each item for who provided it and who typed it:
        i. Form
        ii. Gloss
        iii. SemDom
        iv. Part of Speech
        v. Edits?? (how would this be represented? "Last modified by"? Do we need to know *all* people who edited it, or just the last one?)
        vi. Some indication of whether the data was from:
            1. Before the workshop
            2. During the workshop (words generated, or real-time editing)
            3. After the workshop
3. Ability to work without using the mouse
    a. Rationale: The point of this exercise is to do as much as possible in as short a time as possible. For someone who knows how to use the keyboard commands, those are much faster than using a mouse (or a finicky trackpad), and result in less repetitive stress. There are some mother tongue workers who far prefer the keyboard approach to the mouse.

C. Existing Functionality that has to be retained (Obsolete--replaced by C2 below)

While there are certain new features that justify making a new tool, there are certain features of the existing tool that have to be kept. If these features do not exist in the new tool, it will not be used.

1. Ability to enter words by semantic domain, in response to questions (like Collect Words tool in FLEx)
2. Ability to enter Vernacular Form and Meaning and have them indexed to the Semantic Domain that the current questions are about.
3. Ability to choose whether the Form goes into Lexeme Form or Citation Form, and whether the Meaning goes into Gloss or Definition.
4. Ability to enter additional fields at the same time:
    a. More than one writing system for Form and Meaning.
    b. Grammatical Category
    c. Dialect Label

    d. Scientific Name
    e. Source (person who provided the word). May want to set this “per session”, rather than enter for each word. Will be important on a per-word basis for a paperless workshop.
    f. Etymology: what language the word is borrowed from. (This is a bit problematic because in FLEx this is an entry level field. Needs more specification about how this would be implemented.)
    g. Semantic Domain (view only, for glossing or cleanup phases)
5. Appropriate defaults for entries that might be duplicates (for the interactive mode, or when interactive mode is turned off)
    a. Same Form, different Gloss: one entry, make a new sense
    b. Same Form and Gloss, different SemDom: add SemDom to existing Gloss
    c. Different Form, same Gloss: new entry
    d. When more than one Glossing WS, then “same Gloss” means “all Glosses are identical”
6. Ability to view the existing data in a browse view or detail view, and edit it.
7. Ability to sort and filter the data (by Form, Gloss, SemDom, Dialect(?)), to quickly find anomalies that need fixing.
8. Ability to see/print the words formatted as for a dictionary. Rationale: sheets can be printed daily, sent home with participants for proofing, and corrections entered the next day. It would be okay to have a single hard-coded format, including the fields they have used, rather than allowing the full “Configure Dictionary” functionality in FLEx. [We want to help specify what this single hard-coded format will look like. It should look dictionary-like. But once it is specified for the tool. It doesn’t need to be modified “per workshop”.]
9. Ability to upload to Webonary. Rationale: If participants can see it up on Webonary even after day 1, they become even more excited about the task they are doing, and it gives them ideas for the remaining work.
10. Ability to transfer the data from FLEx before the workshop and to FLEx after the workshop is over. (Each is a one-time operation, not continuous during the workshop. No one will work on the data in FLEx during the time the data is in the new tool.)

C2. Existing functionality (reorganized)

While there are certain new features that justify making a new tool, there are certain features of the existing tool that have to be kept. If these features do not exist in the new tool, it will not be used.

1. Essential for Collect Words task
    a. Ability to enter words by semantic domain, in response to questions (like Collect Words tool in FLEx)
    b. Ability to enter Vernacular Form and Meaning and have them indexed to the Semantic Domain that the current questions are about.

    c. Ability to choose whether the Form goes into Lexeme Form or Citation Form, and whether the Meaning goes into Gloss or Definition.
    d. Appropriate defaults for entries that might be duplicates (for the interactive mode, or when interactive mode is turned off)
        i. Same Form, different Gloss: one entry, make a new sense
        ii. Same Form and Gloss, different SemDom: add SemDom to existing Gloss
        iii. Different Form, same Gloss: new entry
        iv. When more than one Glossing WS, then “same Gloss” means “all Glosses are identical”
    e. Ability to transfer the data from FLEx before the workshop and to FLEx after the workshop is over. (Each is a one-time operation, not continuous during the workshop. No one will work on the data in FLEx during the time the data is in the new tool.)
2. Part of Cleanup Phase (the phase where ‘merge duplicates’ happens; may also happen concurrently with Collect Words phase)
    a. Ability to view the existing data in a browse view or detail view, and edit it.
    b. Ability to sort and filter the data (by Form, Gloss, SemDom, Dialect(?)), to quickly find anomalies that need fixing.
3. Standard part of current RWC workshops, happens during Collect Words phase
    a. Ability to see/print the words formatted as for a dictionary. Rationale: sheets can be printed daily, sent home with participants for proofing, and corrections entered the next day. It would be okay to have a single hard-coded format, including the fields they have used, rather than allowing the full “Configure Dictionary” functionality in FLEx. [We want to help specify what this single hard-coded format will look like. It should look dictionary-like. But once it is specified for the tool. It doesn’t need to be modified “per workshop”.]
    b. Ability to upload to Webonary. Rationale: If participants can see it up on Webonary even after day 1, they become even more excited about the task they are doing, and it gives them ideas for the remaining work.
4. Items that have been entered by various RWC workshops during Collect Words phase (fewer fields than currently in FLEx Collect Words tool)
    a. Ability to enter additional fields at the same time:
        i. More than one writing system for Form and Meaning.
        ii. Grammatical Category (desire ability to enter, not just view)
        iii. Dialect Label
        iv. Scientific Name
        v. Source (person who provided the word). May want to set this “per session”, rather than enter for each word. Will be important on a per-word basis for a paperless workshop.
        vi. Etymology: what language the word is borrowed from. (Not currently in FLEx Collect Words feature, because it is an entry level field. Needs more specification about how this would be implemented.)

vii. Semantic Domain (view, for cleanup phase)

D. Possible Additional Features (Nice-to-have items)

These are features that would enhance the performance of the tool, but they are not required in order for the tool to be used.

1. Specialized cleanup tool that facilitates the specific operations needed for cleanup after a RWC workshop:
    a. Merge entries (keep one or the other, or keep parts of one and parts of the other)
    b. Merge glosses (keep one, concatenate, or edit one or the other)
    c. Merge senses (merge glosses (see b) and keep both SemDoms
    d. Edit SemDoms in a sense: Add, remove
2. Audio recording (and playback) during word collection. [This was an "essential" feature for The Language Conservancy, and it is a "nice to have" feature for some SIL teams.]
3. Possible to use more than one semantic domain list (user can supply their custom list).
    a. Rationale: It is generally recommended that it is better to use a semantic domain list based on the target culture's world view, rather than a list created by an outsider. However, not everyone doing this exercise will have spent enough time in the culture to develop such a list. Thus, the list that ships with FLEx is "better than nothing". It is a stop-gap until they can develop their own, and it also provides a way to compare items across cultures.
4. Possible to collect words based on a wordlist, rather than semantic domain questions.
    a. Functionality: The words in the list become the Gloss, and the user enters the vernacular Form that goes with that Gloss.
5. User can provide their own custom word list.
6. A mode where a typist can just type in words, not linked to any semantic domains or a wordlist.
    a. Rationale: Users say there is not an easy way to do this in FLEx right now--a rapid entry mode with a minimum of fields that they can move through quickly.
    b. Functionality:
        i. Allow user to specify which fields are available (minimum: Form, Meaning, Grammatical Category, Dialect Label, multiple WS for each)
        ii. Fields are easy to see and type in. Tab between fields, Enter accepts a record and moves to the next
        iii. Ability to do it all with the keyboard, no mouse involved, although the mouse should also work.
        iv. This also needs the typeahead functionality, for words already in the database.
        v. Words stay in the order they are typed, until [when?]

vi. This needs the same sorting options as in RWC: Order entered, alphabetical, by length, from end.
vii. Tag the data for the source (person who gave the info, person who typed it)?
viii. A way to indicate the source for a whole batch of data, and then another source for another batch. (Not have to enter the source for each word, but ability to have different sources for different batches.)

E. Dreams for the Future

These are probably not realistic for this project, but it is good to be aware of functionality that we dream of for the future, so current design won't prevent these happening in the future.

1. Tool can detect possible complex forms and suggest what their components might be.
2. Tool has awareness of paradigms and can detect Forms that are not the recommended inflection for a Citation Form.