Writing Micropublications

Michael Cysouw

(Draft from 26 April 2014)

Background

[Introduction] The idea behind micropublications is that in much scholarship there is a growing gap between the sources of information that we have available in our research (books, articles, data collections) and the final use of those sources in our (published) claims and argumentations. Good scholarship of course uses citation to refer to crucial information in a source. However, in many cases the exact formulation as found in the source and the interpretation of this information as used in our own argumentation are not directly connected. There has often been a significant amount of paraphrasing and analysis, making it difficult for a reader to reconstruct how the source-information actually relates to our claims. The more sources we have available, and the more data-driven the research aspires to be, the less amount of space do we have in our own publications to spell out all the interpretative details. This is the gap that micropublications are supposed to fill.

[Digital note taking] In the daily practice of scholarly research, the reading, comprehension and interpretation of sources takes up a large portion of the time (note that this is rather different in many fields of research that collect new data instead of reinterpreting available data). This scholarly hard work often finds its way into excerpts and notes, that used to be collected on filing cards or in notebooks, but which are more and more written down in electronic documents, lying unreachable for outsiders somewhere on a computer hard-drive. It is these files, these excerpts, notes, or filing cards that are micropublications. Too small, too many and often too disorganized to be published in a traditional publication format, but often crucial to understand our stream of thoughts leading from the sources to the final publication. The current proposal is to try and extract these notes from the binary dust in our computers, and share them with our scientific peers. This objective partly arises to show (off) the underlying work that has been ongoing behind a seemingly short paper, but also because these notes are often highly profitably for a colleague to kick-start new research projects. The number and nature of these notes is of course not amenable for traditional print publication, so micropublications will be purely digitally self-published.

[Additional Work?] The preparation of these notes for (micro)publication might seem like additional work, and indeed, a bit of additional work will arise. However, the proposal is to try an minimize the work, while offering a maximum of profitable usage. An important characteristic of micropublications is that they do not have to be stylistically perfect. They might be crudely formulated, grammatically sloppy, maybe even just a few keywords listed together. That should be fine for this context. No need to polish each sentence to perfection. Also on the formal side there should not be a large burden of rules and regulations to be followed. The structure of micropublications should be as free as possible, just a few restrictions on the form are necessary to allow for a future of (automatic) aggregation, distribution and processing.

[Linguistic Typology as use-case] How such a system of (electronic) micropublications might work in practice will be exemplified here for the field of linguistic typology, but I expect that very similar problems (and solutions) will arise in other scholarly fields. In linguistic typology large numbers of descriptions of different languages from all over the world are compared according to one specific linguistic topic. The number of sources used in the course of such a research project can easily be in the hundreds, leading to a situation that the space for discussion in our publications is often reduced to the naming of the source (if at all), with maximally the addition of an approximate page range. That might look like proper scholarly citation practice, but from experience I have to report that it is often very difficult to trace back why somebody concluded from, say, pages 35-39 in a specific grammar about some lesser-known language that this language has, e.g. split ergative case marking (cf. Cysouw 2007).

Practice

[One MP per doculect] Consider a research project with a rough topic (e.g. "the linguistic structure of questions") and dozens, or even hundreds of sources to be investigated. We are going to make one micropublication (MP) for each source. The next step (i.e. the combination, classification, and streamlining of all these MPs) is the part of the traditional publication (with possibly further intermediate steps of MPs that will not be discussed here). The MPs represent the low-level work. Strictly speaking, we are going to make one MP per doculect ("documented lect", cf. Cysouw & Good 2013), i.e. one coherent linguistic entity as described in one specific source. Sometimes there are more than one linguistic entities described in a source, and then we will have to make various MPs. However, in the large majority of cases there is a simple one-to-one match between source and doculect, so it is sufficient to make one MP per source.

[Keep MPs small] Often we might still want to make various MPs for each source. For example, in our project on questions, we might for example be interested in content questions, alternative questions, rhetorical questions, etc. When a source is particularly abundant with information, we might split that into different MPs. Micropublications should be 'micro' after all, so better many different MPs of about 1-2 pages each, than one long excerpt of 35 pages. In that way it becomes easier to later refer to a specific MP as explicating a claim in our published paper. If the interpretation is somewhere hidden in a 35-page mass of information it again becomes impossible for a reader to understand why we make a specific claim in our published work.

[Minimal formal requirements] So we are left with relatively small notes on highly specific topics. I think that we need only two easy formal restrictions on how MPs have to look like. The first formal requirement of MPs is to capture some basic information in the form of a few structured keywords as metadata (i.e. data about the data). However, that will be the only formal content specification of how MPs should be structured. For the rest they can contain any content as deemed necessary. The second formal proposal is to reduce the the number of file-formats for micropublication, and to constrain ourselves what kind of formatting we can use into an MP. These two formal aspects (metadata and file-format) will be discussed in more detail below.

Metadata

The crucial metadata that is necessary for each MP are the following 6 (and only 6!) bits of information:

- **title**

  A name for the MP: short, descriptive, unique. It will become the ID of the MP later on. If your project is small, a language name might be sufficient.

- **creator**

  Your name, possibly also collaborators.

- **date**

  Date, might be automatically added on each save in your workflow. But you can of course just fill it in manually. Try to adhere to standards of date formatting, so this can be read automatically.

- **source**

  Link to the source, typically a link to glottolog.org or any other unique online ID, like a stable OCLC or DOI link.

- **subject**

  Keywords or key-phrases describing the topic of this MP (i.e. the topic of your research). Try to keep them identical across your collection of MPs, and possibly reuse the keywords from others.

- **rights**

  Add some copyright notice. It is strongly preferred to use an open license like CC-0 or CC-by. With CC, please **never** use CC-nd ('non-derivative') for MPs, or any other of your scientific work, as then you officially ban any interesting further use of your work. Even CC-by is not necessary, as the proper citation of your work is enforced by the moral codes of scientific conduct, not by laws or judges.

That's all. Of course you can add more metadata if you want to, but I do not think more is necessary. Two useful optional further fields of metadata might be the following:

- **doculectname**

  Only in case there are multiple languages discussed in one and the same source, add an indication which of those language variant you are discussing. Normally this is not necessary, as the source is a better definition of the language variant than any available naming scheme (cf. Cysouw & Good 2013)

- **relation**

  In case you want to refer to other publications, e.g. other MPs that you agree/disagree with, or want to expand upon, add a complete URL here. Note that this should not be used for regular scholarly reference to authorities (e.g. “according to the definition of ergativity of Dixon 1979”). They should simply be part of the text of an MP in any format you like best.

Except for ‘doculectname’ all of these are basic Dublin Core metadata concepts, which are in widespread use. In practice, I propose to use a YAML-type format at the start of each MP with this information. That would then, for example, simply look like this (note the details of the YAML-format as explained in the link above):

```
---
title:      Tagalog
creator:    Michael Cysouw
date:       26 April 2014
source:     http://glottolog.org/resource/reference/id/117594
subject:
    - content questions
    - HOW
rights:     CC-by
...
```

Just for reasons of readability, the actual reference to the source can be added into the main text, but strictly speaking this is not necessary, as the source is defined by the link in the metadata above.

```
Schachter, Paul and Fe T. Otanes. 1972. Tagalog reference grammar.
Berkeley: University of California Press.
```

Format

[Use simple markup schema] The structure of the rest of the MP is not prescribed: you are free to write what you think is necessary. In contrast to the metadata, this part of the MP is intended for human eyes, not for computers. Still, it would be good to use a form that is easy to process automatically (i.e. easy to convert into different format, easy to share, etc.). As a basic philosophy, you should only use simple markup schemata. In practice I think it is sufficient to use either simple HTML or markdown. Markdown can be used for those who want to immediately write their MPs in simple text format. Please don’t try to write HTML by hand, because you will make errors. However, you can have your documents easily converted to HTML in almost all cases (e.g. from Microsoft Word).

[Use only basic formatting] Further, I would like to propose that there can be only four different kinds of information in MPs, strongly restriction ourselves not to use any fancy layout possibilities:

- humanly-readable text, including simple markup like headings, boldface, italics
- images, including snippets from the source
- lists
- simple rectangular tables (i.e. no merged cells or other strange table-forms).

The whole idea of MPs is, that it should be the thoughts and interpretations, not fancy layout that counts. Try to restrict yourself to easy content-structure, and you will find that it makes life much easier when you do not use all the myriads of options that Microsoft Word allows you to use! Also HTML is much more powerful than just those features, but I propose to restrict ourselves to just these formal concepts. The reason is that the content is then easily saved, archived, converted, and displayed in a variety of manners.

[Problem: tables in markdown] Slightly problematically, basic markdown does not even include tables, but there are many different markdown-dialects that allow for simple tables (and I think linguists will want to have tables for interlinear glossing and for paradigms). For markdown, 'pipe-table' seem to be the most widely used extension of basic markdown. A simple glossed example in markdown might then look like this (the empty lines and the extra line between original and gloss unfortunately seem necessary in most implementations of markdown that I have tested):

Markdown example code:

```

Das|ist|ein|interessanter|Satz.
-|-|-|-|-
DEM|COP|ART|interesting|sentence

"This is an interesting sentence." (p. 44)
```

This then should look somewhat like an interlinear gloss, as shown below. Note that this is not supposed to be the ideal computational format for interlineary-glossed example sentences (it isn't!), but it is a compromise between computer-readability, easy human input, and easy conversion between formats. Further: the details of how this looks on your screen or on paper strongly depend on style descriptions like latex-classes or HTML-CSS. If you don't like the looks, you will have to change these style descriptions, not the MP itself!

Type-setted version:

Das	ist	ein	interessanter	Satz.
DEM	COP	ART	interesting	sentence

"This is an interesting sentence." (p. 44)

[Only rectangular tables] As a general rule: only make the aligned parts of a glossed example into a table (i.e. the first and second line). Do not include the translation into the table structure, as it will mess up the rectangular structure. If you want to add paradigms, also here do not change the basic table structure, i.e. do not try to make merged cells (e.g. to indicate syncretisms). Although that is possible in HTML ("rowspan"), it often leads to wrong results when converting to a different format. So, simply keep tables rectangular, and repeat forms in case of syncretisms/homophonies. For example the German present person inflection in markdown would look like this (this time trying to make the markdown look nicer, which is actually mostly unnecessary work):

Markdown example code:

```
PERSON | SING | PLUR
------ | ---- | ----
1      | -e   | -en
2      | -st  | -en
3      | -t   | -en
```

Type-setted version:

PERSON	SING	PLUR
1	-e	-en
2	-st	-en
3	-t	-en

[Add images from sources] Very important is the possibility to add images, i.e. snippets of the original source. Often it is better not to try and reproduce all the details of an example sentence or a paradigm from the source, but much easier to simply add the page (or part of the page) as an image to the MP. If you are working from a PDF, you can copy part of it into your MP. But you might also consider using your smartphone to photograph a page from a book lying on your desk, and, for example, send it to your computer via email or dropbox. You can of course also completely switch to writing MPs in markdown on a tablet while browsing through the library! The whole idea of making MPs is that it should be easy enough to do on a smartphone or a tablet on the fly, and not just in the complex work-surrounding at your office desk.

Workflow

Everybody has her/his own preferences and habits as to the usage of software in day-to-day practice. Also here I do not want to prescribe any specific workflow. However, a few general guidelines should be considered:

- **save as HTML when not writing markdown**

  If you do not want to write markdown yourself (try it: it's not hard!), you can of course use your favorite word-processor (Microsoft Word, OpenOffice, iWork, etc.), but then you have to save/export your MP as HTML. Microsoft produces terrible, unwieldy HTML, but that is still easier to work with than its proprietary save formats (doc, docx). Consider passing Microsoft HTML through pandoc to turn it into basic HTML that contains the content, but not the details of the style.

- **pack each MP into a directory**:

  Any included images will not be part of the HTML/markdown file (they only contain text), but the images and the content will have to be organized together. The easiest solution is to simply make one directory ('folder') per MP including the file and the images. On conversion to HTML, Microsoft will produce a separate directory with the images, OpenOffice will just list the files individually. Just keep everything as it is, and put it into one directory.

- **use your 'title' from the metadata as the directory name**

  Theoretically this is not important, but in practice it will be easier to find your information.

- **share your work**

  Also in this aspect different people will have different opinions, but I would strongly urge to share your MP-directories in a free and open manner. Consider making a project on github or bitbucket (it's free!) and 'commit' your progress regularly. These websites use **git** for version-control, which means you can easily retrace your steps, and combine your work with others later on. This is one of the reasons to use 'simple' formats (like HTML or markdown), as they work great with git. You might want to take a few hours to try and understand the basics of git, which is highly useful knowledge anyway (and is not normally taught in linguistics classes). In practice you can use easy-to-use software to help you interact with git. As you will mainly be using git for your personal work, you will actually only need 'git commit' and 'git push', not all the other fancy features that git offers for collaborative work.

References

- Cysouw, Michael. 2007. A social layer for typological databases. In Andrea Sansò (ed.) *Language Resources and Linguistic Theory*, 59-66. Milano: Francoangeli.
- Cysouw, Michael & Jeff Good. 2013. Languoid, Doculect, Glossonym: formalizing the notion 'language'. *Language Documentation & Conservation* 7. 331-360.

Note: this paper was written in markdown and typeset into Xelatex using pandoc with minimal style tweaking, using only:

```
pandoc micropublications.md -f markdown -s -o micropublications.pdf
--latex-engine=xelatex -V mainfont="CharisSIL"
```