(Computational) Linguistic Style Guidelines v2.1

Francis BOND
<bond@ieee.org>

February 5, 2013

1 Introduction

Because in linguistics we use words to talk about words, it is important to make clear distinctions between the words being discussed, their meanings and the discussion itself. Fortunately, there is a long tradition of doing this, with some accepted conventions. Unfortunately, they are not always obvious.

These are some notes on style guidelines I made for myself and my colleagues at NTT. Much of the first version is based on (or copied from) the "LSA[1] Style Sheet" LSA (1988), with some additional points from the *Australian Journal of Linguistics* "Notes for Contributors" and the ALS's[2] draft "Guidelines for non-discriminatory usage of language", plus other changes. It differs from the LSA guidelines in that double quotes are used for glosses, following John Lyons (Lyons, 1977, e.g.,); this is useful if your glosses have apostrophes in them, which mine often do. I have updated it with common issues I noticed in my student's writing at Nanyang Technological University, and feedback from colleagues whose writing style I admire.

I have sometimes given suggestions as to how to carry out these guidelines, under the reasonable assumption that you use LaTeX. For more information I recommend Doug Arnold's *LaTeXfor Linguists* `<http://www.essex.ac.uk/linguistics/external/clmt/latex4ling/>`.

If you find any mistakes, or can suggest other useful conventions, or can offer links to even better resources, please contact me.

1.1 Fonts

Avoid using different fonts except for the following purposes:

1. Use *italics* only for cited linguistic forms and for titles of books and journals: The word *grok* is discussed in the *Journal of Martian Studies*. Do not use italics for emphasis, or to mark foreign words used as part of an English sentence: a-priori, ad hoc, inter alia, ipso facto, prima facie, fa con de parler, langue/parole, Sprachgefühl, bunsetsu, etc. – all without italics (LSA, 1988, 3-b).

2. Use ***bold italics*** for lexemes (i.e. ***go*** is the lexeme with the word forms *go, goes, gone, went*) (Huddleston, 1984, xiii).

3. Use SMALL CAPITALS, where it seems essential, to give prominence or emphasis to a word, phrase, or sentence in the text, or to mark a technical term at its first occurrence (LSA, 1988, 3-c). Note that some journals, e.g. *Machine Translation*, prefer **bold** for emphasis. I always mark technical terms with `\txx{}` and define it as either **bold** or SMALL CAPS in the preamble.

[1] Linguistic Society of America

[2] Australian Linguistic Society

4. Use **boldface** for certain forms in Oscan and Umbrian,[3] and when necessary to distinguish Gaulish and other forms originally written in the Greek alphabet(LSA, 1988, 3-d). Sadly, I rarely deal with Oscan and Umbrian so I use it for emphasis (see 3).

5. Use a consistently different font for concepts: `dog`$_1$ is a kind of `animal`$_1$. I like to use **bold sans**, Saeed (2009) uses SMALL CAPS. I recommend marking them with, e.g., `\nd{}` and defining this to be `bold sans` in the preamble:
   `\newcommand{\nd}[1]{\textbf{\textsf{#1}}}`

1.2 Punctuation

1. If the second of a pair of quotes stands at the same point as another mark of punctuation, the quote precedes unless the other mark is itself part of the quoted matter. The word means 'cart', not 'horse'. He writes, 'This is false.' Does that mean 'You heard me'? It means 'Did you hear me?'.

2. Never use quotes to enclose a word or phrase cited as a linguistic example; see Section 1.3 (LSA, 1988, 4-b).

3. Put (double) quotation marks around the titles of shorter works such as journal articles, articles from edited collections, television series episodes, and song titles: "Multimedia Narration: Constructing Possible Worlds"; "The One Where Chandler Can't Cry"; "Let's get Drunk and Funk".

4. Parentheses are used to enclose optional elements: *NTT uses the money (that) it earns* indicates that *that* is optional (Huddleston, 1984, xiii).

5. Square brackets are used to enclose relevant context: *a knowledge [of Japanese]* (Huddleston, 1984, xiii).

6. Footnote markers come after punctuation.[4] This is standard in APA, but seems to surprise my students.[5]

1.3 Cited forms and glosses

1. A letter, word, phrase, or sentence cited as a linguistic example or subject of discussion appears in italics: the suffix *-s*, the word *like*, the construction *mich friert*. Do not use quotation marks for this purpose (LSA, 1988, 6-a).

2. Cited forms and sentences should appear in italics even if the example is set out on a separate line.[6] Examples with morpheme-by-morpheme glosses should be set out as follows[7] (Association for Linguistic Typology, 1997, point 12):

   (1) *manmosu-wa* *zetumetu-shita*
   mammoth-TOP die.out-did
   "Mammoths died out" [jpn]

[3] This recommendation by the LSA makes me unreasonably happy: I am glad to live in a world where some obscure languages get special treatment for no particular reason.

[4] Like this.

[5] `http://owl.english.purdue.edu/owl/resource/560/04/`

[6] Opinion is divided on this. For example, Huddleston (1984) does but Lyons (1977) doesn't. Valued colleagues prefer no italics, I prefer italics. Pick one and be consistent.

[7] This example uses `gb4e.sty` with `\let\eachwordone=\itshape`.

- In particular, left-align examples and glosses word-by-word (not morpheme-by-morpheme); use small capitals for grammatical elements in glosses (whether abbreviated or not), and lower-case for lexical material; underline the portions of the examples which are critical to the argument;
- Use hyphens to separate corresponding components in examples and glosses; use full stops for components in glosses that are (for whatever reason) unseparated in the example; and use (and explain) whatever further boundary markers and glossing conventions you may find useful, especially when correspondences between examples and glosses are more intricate;
- It is often useful to tag each IGT instance with the language it comes, using ISO 693-3 code for that language at the end of the translation line, in square brackets;
- Exactly how much you show in the glosses depends on what you are doing. For example, an alternative gloss of (1) might explicitly label the past verb form. You do not need to do this if the details of the morphology are not relevant to the general discussion.

  (2) *manmosu-wa* *zetumetu-shita*
  mammoth-TOP die.out-do.PFV

  "Mammoths died out" [jpn]

  However, it is important to keep the same number of tokens and hyphens in the source and gloss line.
- Put the final punctuation mark of the translation inside the quotation mark; whether you prefer to begin example sentences with upper or lower case letters and to punctuate them at the end or not (while example words or phrases should not be capitalized nor punctuated), be consistent in your preferences; carefully acknowledge (primary) sources of examples, and avoid quoting any at second hand;
- Refer to numbered examples (and other such numbered items) in the running text by their number and possibly letter within parentheses (e.g. "as seen in (12) and (17a-b), number in Mohawk ..."), but avoid beginning a sentence with such a parenthesized reference number.
- See the Leipzig glossing rules (Comrie et al., 2008) for more guidelines on interlinear glossing: `<http://www.eva.mpg.de/lingua/resources/glossing-rules.php>`
- You may also add the original orthographic form on the first line

(3) マンモスは 絶滅 した
manmosu-ha *zetumetu-shita*
mammoth-TOP died out

"Mammoths died out"

3. Cited forms may also appear in phonetic or phonemic transcription, enclosed in square brackets or in slant lines: the suffix [s], the word /layk/. Symbols between brackets or slants are never underscored or put into different fonts (LSA, 1988, 6-b).

4. Forms in a language not written with the Latin alphabet must be transliterated (or transcribed), unless there is a cogent reason for citing them in the original characters. This provision applies to Greek as to other languages (LSA, 1988, 6-c).

5. In particular, for Japanese, a standard phonetic transcription such as Hepburn romanization should be used. For example: the topic and object markers should be written as

-wa and *-o* respectively; long vowels should be either doubled (*toozen*) or marked with an accent (*tōzen*); and *shi/si, ji/zi chi/ti tsu/tu ...* should be used consistently (Anonymous reviewers TMI'95).

6. Cited forms in a foreign language should be followed at their first occurrence by a gloss in double quotation marks. No comma separates the gloss from the cited form: Japanese *hitsuji* "sheep" is a noun. No comma follows the gloss unless it is required by the sentence as a whole: Latin *ovis* "sheep", *equus* "horse", and *canis* "dog" are nouns. Note that the punctuation follows the quote.

7. If you include examples in the original language, you still need the transliteration and the gloss. For example, 羊 *hitsuji* "sheep" is a noun, 美しい *utsukushii* "beautiful" is an adjective.

1.4 Bibliographic reference

1. There are many different styles for bibliographic references. I like to use the unified linguistic style guide, but also give here some notes from the Lingusitic Society of America. The American Psychology Association (APA) has a nice consistent style but recommends the use of initials for authors first names, which is not a good thing.

2. The Committee of Editors of Linguistics Journals has developed a unified style sheet for linguistics journals (CLExJ, 2007). I strongly recommend it. You can find it here: `http://celxj.org/`. It is used in this style guide.

3. Full citation of literature referred to should be given in a bibliography at the end of each article or review. Within the text, brief citation will be made, normally by giving the author's surname, the year of publication, and the page number(s) where relevant. Such brief citations should be given in the body of the text, not in footnotes, unless they refer specifically to a statement made in a footnote (LSA, 1988, 9-a).

4. The brief citations given in the text should take such forms as '(Bloomfield, 1933)' or '(Hocket, 1964:240–241)'. Note that the page numbers given here are only for the passage to which reference is made, not for the whole paper. Use initials for authors' given names only when necessary to distinguish, e.g. N. Chomsky and C. Chomsky within a single article. If the author's name is part of the text, use this form: 'Bloomfield (1933:264) introduced the term ... ' (LSA, 1988, 9-c).

5. Where the names of authors or editors appear in the list of references, do not replace given names with initials, unless such abbreviation is the normal practice of the individual concerned: thus Miller, Roy Andrew (not Roy A. or R. A.), Hooper, Joan B. (not J.B. or J.); but Palmer, F. R. (LSA, 1988, 9-d). Indeed, to make it easier to track down authors, it is best to use their full names as far as possible. More specifically, use the full form of the name **as listed in the paper**, so that if the authors initialize their own names, use that, and if they like to use an initial (such as *J. Ross Quinlan*) preserve this.

6. Using the `natbib` style file the following styles of quotation are made possible:

   citep[pageno]{key} Which produces citations with both author and year, enclosed in parenthesis: (Huddleston, 1984, 111).

   citet[pageno]{key} Which produces citations with the author followed by the year enclosed in parenthesis: Huddleston (1984, 111).

   citealt[111]{key} Which produces citations with the author followed by the year, no parenthesis: Huddleston 1984, 111.

   citeauthor{key} Which produces the author only, no parenthesis: Huddleston.

   citeyear{key} Which produces the year only, no parenthesis: 1984.
   `natbib` has many other useful options, see `http://www.ctan.org/pkg/natbib` for the full documentation

7. Reference citations for two or more works within the same parentheses: List two or more works by different authors who are cited within the same parentheses chronologically. Separate the citations with semicolons: Several studies (Balda, 1980; Kamil, 1988; Pepperberg & Funk, 1990).

   Exception: You may separate a major citation from other citations within parentheses by inserting a phrase such as see also, before the first of the remaining citations, which should be in chronological order: (Minor, 2001; see also Adams, 1999; Storandt, 1997).

1.5 Abbreviations

1. Abbreviations ending with a letter in lower case have a following period; abbreviations ending with a letter in upper case have none (LSA, 1988, 7-a).

2. Names of languages prefixed as adjectives to linguistic forms are often abbreviated: E or Eng., ME, OE, J or Jap., Ger., Fr., OFr., Gk. (not Gr.), Skt. (not Skr.), IE (not I-E), PIE. But names of languages used as nouns are not abbreviated: the meaning of OE *guma*, the meaning of *guma* in Old English (LSA, 1988, 7-b).

3. Titles of well-known journals are often abbreviated in bibliographical references: AA, IJAL, BSOAS. The regular abbreviation of *LANGUAGE* is Lg (LSA, 1988, 7-c).

4. Abbreviate grammatical terms directly attached to linguistic forms: Latin inf. *portāre*, 1sg. pres. ind. *porto*, 2 pl. *portātis*, 3sg. impf. *portābat*. But do not abbreviate such terms in other uses: the Latin imperfect in *bā* (LSA, 1988, 7-d).

5. Explain abbreviations of technical terms (unless they are self-explanatory), and list them in a separate (numbered) note when they are numerous; limit their use to category labels in glosses, tables, and figures and to formulae, and do not let abbreviations interrupt the smooth flow of your prose (Association for Linguistic Typology, 1997, point 9).

1.6 Figures and Tables

1. Plan each table so that it will fit into the printed page without crowding. Leave ample white space between columns. Do not use vertical and horizontal rules unless the table would be unclear without them. (LSA, 1988, 10-a)

System	# features	Accuracy	Coverage	F-measure
Baseline	20	0.5	**0.9**	0.62
Them	20,000	0.6	0.7	0.64
Us	2,000	**0.9**	0.6	**0.72**

Table 1: Good Table

System	# features	Accuracy	Coverage	F-measure
Baseline	20	0.5	0.9	0.615384
Them	20000	0.6	0.7	0.646154
Us	2000	0.9	0.6	0.72

Table 2: Bad Table

2. Column heads should be short, so as to stand clearly above the several columns. If you need longer headings, represent them by numbers or capital letters and explain these in the text preceding the table (LSA, 1988, 10-b).

3. If two or more tables appear in one article, number them and refer to them by number. Do not speak of the 'preceding' or 'the following table'; the printer may not be able to preserve its original position (LSA, 1988, 10-c).

4. Each figure or table should have a caption below it. The caption contains the number and optionally a concise title, sometimes also (as a separate line) a brief explanation or comment (LSA, 1988, 10-d). The caption should have only the initial letter capitalized.

5. Columns of text should typically be left justified, numbers should be right justified (No citation, just believe me).

6. It is the convention in computational linguistics that the best result in a column (or row) is made bold: this makes it easy to pick out. If you do this, you should try to use a font where the width of bold numbers is the same as normal font (I like to use Palatino with the `mathpazo` package).

7. Try to make sure that the ratio of $\frac{\text{data ink (the part with content)}}{\text{total ink}}$ is as high as possible: get rid of unnecessary lines in tables and graphs (Tufte, 2001, p93).

8. Refer to a table or figure as "Table/Figure X", etc., with the first letter capitalized.

9. Compare the good table (Table 1) with the bad table (Table 2). The bad table has: too many lines, too many decimal places, left aligned numbers in one column, no comma after thousands, the F1 results don't line up on the decimal point (and show spurious precision), and the best results are not helpfully made bold.

1.7 Discriminatory use of language

Be sensitive to the social implications of language choice. In particular seek wording free of discriminatory overtones wherever possible.

1. Avoid so-called masculine generics with sex-indefinite antecedents or *man* and its compounds except in unambiguous reference to males. For example: *Everybody has their own preference*. (Instead of: *Everybody has his own preference*.)

2. Avoid adding modifiers or suffixes to nouns to mark the sex of referents unnecessarily. Such usage promotes sexual stereotyping in two ways.

(a) By highlighting the referent sex, modification can signal a general presupposition that referents will be of the other sex (*lady professor, male secretary*) and thus that these referents are aberrant.

(b) Conventionalized gender marking 'naturalises' the presumptive or unmarked sex of the noun's referent (*stewardess, cleaning lady* as opposed to *steward, cleaner*).

3. Use parallel forms of reference for women and men. For example Mr Ogura and Ms Kamezaki, not Kentarou and Ms Kamezaki.

4. Avoid explicitly religious expressions: *. . . witness 'donkey' pronouns, so christened because of . . .* (Instead: *. . . witness 'donkey' pronouns, given that name because of . . .*).

5. Avoid stereotyping or demeaning characterizations and derogatory content in examples.

6. In glossing forms from another language, do not introduce gender-specificity or asymmetry absent in the original.

7. Use gender neutral names in examples: for example *Kim smiled* rather than *Mary smiled*. Especially avoid using male names as agents and female names as patients.[8]

1.8 What NOT to say

This is based largely on Green & Morgan (1996)[9] which goes into a lot more detail. Your claims will be more testable, and your prose more persuasive if the following kinds of expressions **do not** appear in it.

Hedges

should, may, can, some, seem, considered, likely, scare quotes.

Self-congratulatory evaluative terms and intensifiers

simply, easily, strongly, clearly, naturally, intuitive(ly), obviously, certainly, quite, rather, very

When linguists write *It is clear that . . .* or *Obviously, . . .* it is often the case that they really believe that it (whatever it is) is not clear or obvious; if it were, why would they feel compelled to try to persuade you that it **is** clear or obvious? Let the assertion stand on its own instead of drawing attention to your insecurity by using these cheap cosmetics. If it can't stand on its own, you need to spell out your argument in more detail, or maybe come up with a better claim, so that it can stand on its own.

Unnecessarily personalizing expressions

I submit, I maintain, I claim, I propose

Green & Morgan (1996) says you should avoid the first person in summaries (no cheating with passives!). Using the third person forces you to talk about the relation of ideas to each other, not your relation to the ideas. I [Francis Bond] think we should use the first person in descriptions of experiments. I think style is gradually changing here.

[8] The use of gender neutral names is standard practice in the HPSG community, citation needed.

[9] And the sadly vanished `http://lees.cogsci.uiuc.edu:80/~green/wp/stoplist.html`.

Saying what you mean

Whatever you're working on, when you arrive at a concise and precise representation of the claim that you are defending, **do not "cut-and-paste" to represent it in exactly that wonderful way every time you refer to it.** Do NOT repeat all the details every time you refer to it, especially when those details are irrelevant to the part of the hypothesis that is at issue in the particular discussion. When you present the hypothesis with exactly the same words every time you mention it, you do a lot of damage to yourself.

- The representation becomes a **litany**–something you say without thinking about it without meaning it. This removes an opportunity to think about what you really mean every time you say it.
- It bores people, and makes them fall asleep. This is not a very effective way of convincing them of your point or that you are brilliant.
- It implies that you don't know any other way to say what you mean, and that implies that you don't really understand what you are saying.

References

Association for Linguistic Typology. 1997. Instructions for contributors. `http://148.88.14.7/alt/stylesh.htm`.

CLExJ. 2007. Unified style sheet for linguistics. `http://celxj.org/downloads/UnifiedStyleSheet.pdf`. (accesed 203-01-28).

Comrie, Bernard, Martin Haspelmath & Balthasar Bickel. 2008. Leipzig glossing rules. `http://www.eva.mpg.de/lingua/resources/glossing-rules.php`. (accesed 2010-01-30).

Green, Georgia M. & Jerry L. Morgan. 1996. *Practical guide for syntactic analysis*. Stanford: CSLI Publications, Stanford: CSLI.

Huddleston, Rodney. 1984. *Introduction to the grammar of English* Cambridge textbooks in linguistics. Cambridge: Cambridge University Press.

LSA. 1988. LSA style sheet. `http://www.umich.edu/~archive/linguistics/LSA.style.sheet`. (accesed 2010-01-30).

Lyons, John. 1977. *Semantics*. Cambridge: Cambridge University Press.

Saeed, John I. 2009. *Semantics*. Wiley-Blackwell 3rd edn.

Tufte, Edward R. 2001. *The visual display of graphical information*. Graphics Press 2nd edn.

Revision History

v2.1 Incorporating comments from Emily, Michael, Tim and Zina, put online (2013)

v2.0 Revised for use at NTU (2011)

v1.0 Created at NTT (1997?)