

e.g.

David J. Weber  
Summer Institute of Linguistics  
(david\_weber@sil.org)  
Workshop on Web-Based Language Documentation and Description  
12-15 December 2000, Philadelphia, USA

20th December 2000

#### Abstract

Good language descriptions liberally illustrate their claims with examples, ideally drawn from natural discourses of diverse genre. Typically, an example is a text fragment enriched in various ways:

- ★ Linguistic: morpheme boundaries, morpheme glosses, free translation, punctuation, brackets to indicate structure, categories, subscripted indices, grammaticality judgements (\*, ?), empty categories (PRO, pro...), and so forth.
- ★ Attention-directing: italics, bolding, underlining or other highlighting mechanism; multiple examples may be conflated by means of braces or parentheses to isolate parts that differ, and so forth.
- ★ Identification: numbering relative to other examples, example-internal identifiers (like a., b.), reference to the origin of the example (in some other document), and so forth.
- ★ Contextual: speaker's name, age, sex, dialect; the context of use; the register; the nature of production (written, oral, recorded), and so forth.

Such elements are combined and given visual form.

For linguists to provide good, web-based language descriptions they must be able to create examples with diverse enrichments and layouts, ideally linked to both the corpus from which the fragment is drawn and the descriptive contexts in which it is displayed. They need hospitable authoring environments with tools that are powerful and flexible, yet reasonably easy to learn and use.

The purpose of this paper is to initiate discussion on what needs to be in place for linguists to provide examples in their web-based descriptions of human languages.

## Contents

<b>1 INTRODUCTION</b>	<b>1</b>
<b>2 THE CHALLENGE OF RENDERING</b>	<b>2</b>
2.1 The layout of the major elements . . . . .	3
2.2 Conflation . . . . .	4
2.2.1 Braces . . . . .	4
2.2.2 Brackets . . . . .	5
2.2.3 In-line conflation . . . . .	6
2.3 Line wrapping and page breaks . . . . .	6
2.4 Cyber-effects . . . . .	7
<b>3 A POSSIBLE FRAMEWORK</b>	<b>7</b>

## 1 INTRODUCTION

The linguistic literature is populated by a menagerie of “specials”: tables, trees, maps, lists of various sorts, . . . HPSG’s rectilinear attribute-value matrices and RG’s curvaceous stratal diagrams. Among these, the most important for baseline language descriptions is THE EXAMPLE.

Good language descriptions liberally illustrate their claims with examples. (My grammar of Huallaga Quechua, for example, has over 1700.) Ideally, these are drawn from a collection of natural discourses of diverse genre. Many are isolated sentences elicited from a native speaker of the language. And some are non-examples: ungrammatical (unacceptable) sentences.

Typically, an example is created by copying a text fragment into the description and enriching it in various ways. Some enrichments are strictly linguistic:

- ★ The boundaries between morphemes are marked by hyphens.
- ★ Each morpheme is identified by a gloss (“tag”).
- ★ A free translation is included, usually in some major language.
- ★ Structural units may be indicated with brackets.
- ★ Categories may be indicated, usually by subscripts to brackets.
- ★ Functions (e.g., subject, instrument, source,...) may be given for noun phrases.
- ★ Grammaticality judgements may be indicated (\*, \*\*, ?, ??).
- ★ Subscripted indices may indicate coreference or disjoint reference between referring expressions.
- ★ Empty categories may be indicated: e, t, PRO, pro,...
- ★ Ellipsis marking may indicate that the example, as presented, is only part of a larger unit.
- ★ Explanatory notes may be attached to morphemes or larger units.

And so forth.

Some enrichments are aids to the reader:

- ★ Attention may be directed to a particular aspect by italics, bolding, underlining, or some other highlighting mechanism.
- ★ Multiple examples may be conflated by means of braces or parentheses to facilitate the comparison between alternatives.
- ★ Punctuation may be added.

And so forth.

Some enrichments serve to identify the example in the context of the document in which it is cited:

- ★ Normally each example bears a number identifying it in the description. This number is used for referring to the example, either from within the description or from some other document.
- ★ An example may employ internal identifiers for alternatives within the example itself; e.g., “See example 32.c.” might direct the reader to the third alternative within example 32.

And so forth.

Some enrichments give information about the example’s origin:

- ★ Speaker: name, age, sex, dialect,...
- ★ Context: when, where, to whom... the sentence was spoken
- ★ Register: formal, colloquial,...
- ★ Nature of production: written, oral, recorded. . .
- ★ Textual context: a reference to the text from which the example is drawn, and its position therein
- ★ Residence: the document, archive, collection,... to which the example (fragment) belongs?

And so forth.

Elements of these various types are combined and given visual form, traditionally on a printed page but increasingly on a computer monitor.

There are conventions for how they are combined, but to my knowledge only loose, unwritten ones. Thus, part of the challenge for the enterprise of providing web-based tools for handling examples is to develop guidelines leading to good practice and curbing individuals’ tendencies toward the idiosyncratic. We need a style sheet for examples!

## **2 THE CHALLENGE OF RENDERING**

Rendering an example involves laying out the major parts, sometimes using braces, parentheses or brackets to conflate alternatives, deciding how to wrap long lines or where break examples over a page, and directing the reader’s attention with various highlighting mechanisms. We discuss these in turn.

## 2.1 The layout of the major elements

A rather standard form of example has a number (NUMBER); the text fragment, with hyphens dividing morphemes (MORPHEMES); morpheme glosses, aligned either morpheme-by-morpheme or, more commonly, word by word (GLOSSES); and a free translation (TRANSLATION). These are laid out as follows:

(NUMBER) MORPHEMES  
GLOSSES  
'TRANSLATION'

For example:

- (1) [[yapya-y]-ta uša-na-n]-ta-ši šuya-ra-yka-n  
plow-INF-OBJ finish-SUB-3P-OBJ-RPT wait-DUR-IMPV-3  
'He is waiting for him to finish plowing.'

There are various reasons for also including the example written in the practical orthography:

- ★ Linguistically-oriented representations (phonetic, phonemic, morphophonemic) may be inaccessible to speakers of the language whereas including the traditional/conventional writing system may make it easy for them to read. (By the way, except for phonology papers, English examples use the practical orthography; indeed, most readers would be very put off if they had to read English examples in a phonetic, phonemic, or morphophonemic representation.)
- ★ Linguists who seriously study a language should learn its writing system so as to be able to benefit from other documents written in the language. This learning can occur most painlessly by seeing the practical orthography side by side with a more linguistically-oriented representation.
- ★ The practical orthography may have information not contained in the morphemic representation. For example, in example 5 below the practical orthography represents phonetic detail not indicated by the morphemic form.

The most normal place for the practical orthography (WRITING) is perhaps on the very first line. (Because it is primarily for readers who can read and understand it, it is not necessary to align it with the glosses.)

(NUMBER) WRITING  
MORPHEMES  
GLOSSES  
'TRANSLATION'

For example:

- (2) "¡Ama aywaychu!" nir willashcă.  
ama aywa-y-ču ni-r wiła-ška-:  
no go-2IMP-NEG say-SS advise-PRF-1  
'I told him not to go. (lit. I advised him saying "Do not go!")'

When examples are short, this sort of layout may waste space: when printed, it can increase the cost; when viewed on-line, it can push relevant text off the screen. Therefore, when space permits it may be desirable to use alternative layouts. The translation, for example, might follow the morphemic representation:

(NUMBER) WRITING  
MORPHEMES 'TRANSLATION'  
GLOSSES

For example:

- (3) Liguitya yachachimanga.  
ligi-y-ta yača-či-ma-nqa 'He will teach me to read.'  
read-INF-OBJ learn-CAUS-⇒1-3FUT

Or the written form may also fit there:

(NUMBER) MORPHEMES WRITING 'TRANSLATION'  
GLOSSES

For example:

- (4) huk runa ka-ša Juc runa casha. 'There was a man.'  
one man be-3PRF

And short examples might all fit on one line:

(NUMBER) MORPHEMES (GLOSSES) WRITING ‘TRANSLATION’

For example:

- (5) a. imana-ša-taq (what.do-3PRF-¿?) ¿Imanashataj? ‘What did he do?’  
 b. imana-šaq-taq (what.do-1FUT-¿?) ¿Imanashātaj? ‘What will I do?’

If—as I am assuming—the layout of the major elements depends on the available space, and if in a web-based environment column width is under the control of the reader, then the rendering engine must include a component that adjusts the layout depending on the available space and user preferences.

## 2.2 Conflation

Two or more examples may be conflated by means of braces, parentheses or brackets. In some cases only words are conflated; in others the morphemes within a word might be conflated. For example, consider the following, taken from the *International Journal of American Linguistics* 65:159:

(44a) li:-ta-pa:-chi’:-ní:t tasiw caja  
 INSTR-INGR-belly-tie-PFV rope box  
 ‘The box has been tied up with a rope.’

(44b) li:-ta-maq-chi’:-ní:t tasiw caja  
 INSTR-INGR-body-tie-PFV rope box  
 ‘The box has been tied up with a rope.’

These could have been conflated as follows:

(44) li:-ta-  $\left. \begin{array}{l} \text{a. pa:} \\ \text{belly} \\ \text{b. maq} \\ \text{body} \end{array} \right\} \text{-chi’:-ní:t tasiw caja}$   
 INSTR-INGR  $\left. \begin{array}{l} \text{-tie-Pfv} \\ \text{rope box} \end{array} \right\}$   
 ‘The box has been tied up with a rope.’

There are various reasons for conflating examples:

- \* It makes the example more readable. Without conflation the reader must scan the examples to isolate the parts being compared/contrasted. With conflation this is immediately obvious.
- \* It makes better use of space. Thus, for a printed page it is more economical, and for a computer monitor it allows more context to be kept in view.
- \* It may simplify wrapping examples across lines and breaking lines over pages: Without conflation, two or more parallel lines normally wrap or break independently, which means the layout becomes increasingly difficult to read as the column is narrowed. When two or more examples are conflated to a single line, this is more likely to wrap or break without creating problems (assuming that the portion in braces moves as a piece).

### 2.2.1 Braces

*Linguistic Inquiry* is now virtually devoid of braces except the *characters* { and }, probably as the result of making the journal available on-line with HTML technology. This has a cost; for example, consider the difficulty of reading and the wasted space in 6, *LI* 30:658, example 33. (“BP” stands for Brazilian Portuguese.)

- (6) a. Eu encontrei as minhas velhas amigas e amigos (BP)  
 I met the.F.PL my.F.PL old.F.PL friends.F.PL and friends.M.PL  
 juntos.  
 together.M.PL
- b. Eu encontrei as minhas velhas amigas e amigos  
 I met the.F.PL my.F.PL old.F.PL friends.F.PL and friends.M.PL  
 no mesomo dia.  
 on.the same day.  
 ‘I met my famous old female friends and male friends together/on the same day.’

This could be conflated as follows:

- (7) Eu encontrei as minhas velhas amigas e amigos (BP)  
 I met the.F.PL my.F.PL old.F.PL friends.F.PL and friends.M.PL
- $$\left\{ \begin{array}{l} \text{a. juntos.} \\ \text{together.M.PL} \\ \text{b. no mesomo dia.} \\ \text{on.the same day.} \end{array} \right\} \text{ 'I met my famous old female friends} \\ \text{and male friends } \left\{ \begin{array}{l} \text{a. together} \\ \text{b. on the same day} \end{array} \right\} \text{.'}$$

There follow further examples (from Huallaga Quechua) to illustrate some of the complexities that examples may contain. Example 8 illustrates indices:

- (8) a. Magarcamaptin jaytashurayqui.
- $$\text{maqa-rkU-ma-} \left\{ \begin{array}{l} \text{a. -pti} \\ \text{DS} \\ \text{b. *-\špa} \\ \text{SS} \end{array} \right\} \text{-n}_i \text{ hayta-šu-ra-yki}_j$$
- hit-UP- $\Rightarrow$ 1 3P kick- $\Rightarrow$ 2-PST-2P
- a. 'After he<sub>i</sub> hit me, he<sub>j</sub> kicked you.'

There may be braces within braces, that is, conflation within conflation:

- (9) b. Magarcushpan jaytamaran.  
 c. Magarcur jaytamaran.
- $$\text{maqa-rku-} \left\{ \begin{array}{l} \text{-špa} \\ \text{SS} \end{array} \left\{ \begin{array}{l} \text{a. *-\emptyset} \\ \text{b. -n (-3P)} \end{array} \right\} \right\} \text{hayta-ma-ra-n}$$
- $$\text{hit-ARR} \left\{ \begin{array}{l} \text{-r} \\ \text{SS} \end{array} \left\{ \begin{array}{l} \text{c. -\emptyset} \\ \text{d. *-\nin (-3P)} \end{array} \right\} \right\} \text{kick-}\Rightarrow\text{1-PST-3}$$
- b,c. 'After he<sub>i</sub> hit him<sub>j</sub>, he<sub>i</sub> kicked me.'

Note that in 9 there are no (apparent) right braces to match the smaller left braces. A conflation tool should allow for this possibility, but the challenge will be to characterize the environments in which a matching delimiter can be suppressed. And perhaps cases like that illustrated in 10, where a single left brace appears to be matched by two right braces, should be disallowed. (In fact, one large right brace and two small left braces have been visually suppressed.)

- (10) pay-ta rika- }  
 him-OBJ see }  

$$\left\{ \begin{array}{l} \text{a. -na:-paq} \\ \text{-SUB-1P-PUR} \\ \text{b. -q} \\ \text{-SUB} \\ \text{c. -na:-paq} \\ \text{-SUB-1P-PUR} \\ \text{d. *-q} \\ \text{-SUB} \end{array} \right\} \left\{ \begin{array}{l} \text{\šamu-ška-:} \\ \text{come-prf-1} \\ \text{\šuya-ra-yka-ška-:} \\ \text{wait-DUR-IMPV-PRF-1} \end{array} \right\}$$

a,b. 'I came to see him.'  
 c. 'I was waiting to see him.'

Example 11 illustrates the inclusion of contextual information:

- (11) qam-pis maqa-ma-ška-nki- }  
 you-ALSO hit- $\Rightarrow$ 1-PRF-2 }  

$$\left\{ \begin{array}{l} \text{a. -mi (-DIR)} \\ \text{b. -ši (-RPT)} \\ \text{c. -či (-CNJ)} \end{array} \right\} \text{ 'You also hit me.'}$$

a. SITUATION: I felt you hit me and realized it was you.  
 b. SITUATION: I was drunk when you hit me, but someone told me that you had hit me.  
 c. SITUATION: Various people hit me and I surmise that you were one of them.

It would take scores of examples to illustrate the range of possibilities but hopefully these suffice to suggest the sorts of complexities that need to be addressed.

### 2.2.2 Brackets

Square brackets are sometimes used in contrast to (curly) braces to signal a correspondence among bracketed elements. For example, consider 12:

- (12) a. **kay**-man aywa-**mu**-n ‘He comes here.’  
 here-GOAL go-TO.HERE-3  
 b. **čay**-man aywa-n ‘He goes there.’  
 there-GOAL go-3

This might be conflated as in 13, where *kay* corresponds to *-mu* and ‘*He comes here.*’, while *čay* corresponds to the absence of *-mu* and ‘*He goes there.*’:

$$(13) \begin{bmatrix} \text{kay} \\ \text{here} \\ \text{čay} \\ \text{there} \end{bmatrix} \begin{matrix} \text{-man} & \text{aywa-} \\ \text{GOAL} & \text{go} \end{matrix} \begin{bmatrix} \text{-mu} \\ \text{TO.HERE} \\ \text{-}\emptyset \end{bmatrix} \begin{matrix} \text{-n} \\ 3 \end{matrix} \begin{bmatrix} \text{‘He comes here.’} \\ \text{‘He goes there.’} \end{bmatrix}$$

Square brackets seem to be used less and less (although perhaps they are still used in phonology). Much as one might wish this convention to be a thing of the past, if linguistic articles from decades past are to be formatted in the web-based framework, it will be necessary to support the bracket convention.

### 2.2.3 In-line conflation

There are in-line conflations:

$$\star [X A/B/\dots Y] \text{ is equivalent to } X \begin{Bmatrix} A \\ B \\ \vdots \end{Bmatrix} Y, \text{ which conflates } \begin{Bmatrix} X A Y \\ X B Y \\ \vdots \end{Bmatrix}.$$

The examples in 14 are from *LI* 30:545:

- (14) If/As/When you eat more, you want correspondingly less.  
 If/\*As you had eaten more, you would want less.

$$\star [X (*A) Y] \text{ is equivalent to } X \begin{Bmatrix} \emptyset \\ *A \end{Bmatrix} Y. \text{ Example 15 is from } LI \text{ 30:568. (By the way, “t” is a trace, not a typographical error.)}$$

- (15) This is the kind of rice that the quicker (\*that) you cook t, the better it tastes.

$$\star [X *(A) Y] \text{ is equivalent to } X \begin{Bmatrix} *\emptyset \\ A \end{Bmatrix} Y.$$

As argued regarding layout (section 2.1), the rendering engine must include a component that conflates examples based on the line length and user preferences. This must be under the control of the author of the description, who may give certain options to the reader. For example, a reader may wish to “deconflate” alternatives, seeing them as a list of sentences without braces, brackets or parentheses.

## 2.3 Line wrapping and page breaks

When an example must be broken across a page boundary, it is important that this be done at certain points and not at others. For example, the glosses should never be separated from the morphemes to which they correspond.

Likewise, when an example is too long to fit on a single line, it must be “wrapped” in a way that does not interpose text between, say, the morpheme decomposition and the corresponding glosses.

To break some lines attractively may require hyphenation. For example, for the Spanish version of my Huallaga grammar both the Quechua written form (practical orthography) and the Spanish translation were hyphenated, that is, “discretionary hyphens” were computationally introduced. This process differs from language to language subject to convention, syllable structure,... perhaps even taste (a subjective aesthetic criterion).

Although quite obvious, we should not forget that the space in which an example is rendered depends on the document context: if it is embedded within an item in a list, where each item is indented, then the effective column width for the example is correspondingly narrower.

## 2.4 Cyber-effects

Much could be done in a web-based environment that could not be done in the ink-on-paper context:

**inspect the context:** Traditionally, what you see is all you get: although an example might be a fragment of a text, it is not possible to see the preceding or following text. In the future, when examples are enriched fragments of online texts, software should allow the user to dynamically inspect the text surrounding the example.

**toggle on/off parts:** It may be useful to turn on or off the display of certain kinds of information. For example, native speakers may wish to toggle off parts they do not need, such as the morphemic representation, glosses and translation. Linguists not familiar with the language may wish to toggle off the practical orthography, while linguists familiar with the language may wish to suppress the gloss. And so forth. User should be able to tailor the display of information to meet their needs and preferences.

**buttons and hot zones:** Buttons could be provided to activate certain kinds of secondary information, e.g., the speaker's biographical information, the context of use, the example's "residence," and so forth. Perhaps if the gloss is toggled off, morphemes could be "hot," so that clicking on them would show information about the morpheme: the gloss, the category, perhaps even taking the user to a database entry about that morpheme.

**enhanced focus mechanisms:** Traditionally attention is directed by static effects like bold or italic type, or by underlining. Now it would be possible to use coloring, and effects like blinking. It might be useful to have three variants of comparison, one to signal 'note the similarity of these', one to signal 'note difference between these', and a default for simple comparison.

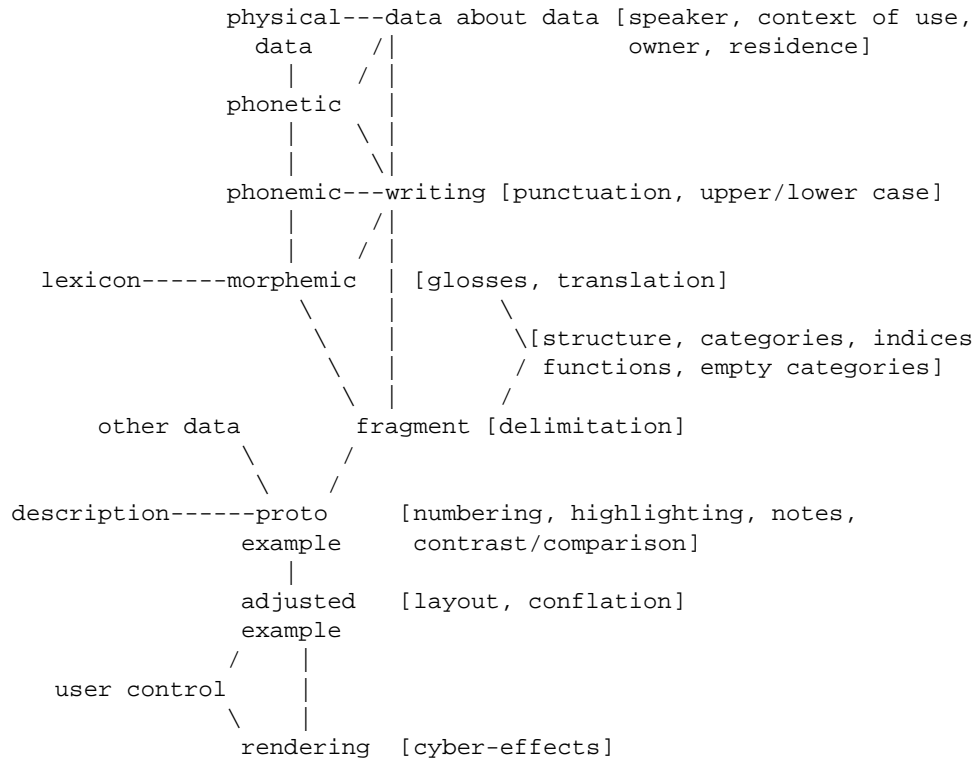
## 3 A POSSIBLE FRAMEWORK

Here I make some very tentative remarks about a possible framework for handling examples. The primary suggestion is that a rendered example be produced by a progression of enrichments starting from a text fragment.

As a starting point there may be physical data, like an acoustic recording, or a phonetic transcription. As the end point, the example is displayed—in the context of a description—on a terminal or printed on a page. Between the data and display there is a network of linked "points" at which information is drawn from other points and enriched as specified by an author.

Here is a very rough sketch. The items in square brackets suggest the types of enrichments that might be made at a point. Some notes follow.

Table 1:



- ★ The “data about data” might link to various points, such as to the written text, the morphemic representation,...
- ★ Writing has punctuation, case, context sensitive renderings,... It bears a close relationship to the phonetic, phonemic, or morphemic representations but is usually not isomorphic to any of them. A general approach would be to treat the written form as a separate point (data object) with links to the phonetic, phonemic, or morphemic representations. The nature of those links would depend on the nature of the writing system.
- ★ Morpheme glosses (possibly in multiple languages) must be linked to the morphemes in the morphemic representation. If a lexicon is available, the glosses should be stored there and acquired by the morphemic representation through links to the lexicon.
- ★ Structure, categories, functions, empty categories, indices and any other enrichments directly tied to the text itself might be given with the morphemic representation or might be given with the proto example. (The latter would allow different authors to represent different analyses for the same fragment.)
- ★ “Other data” is included for (1) elicited sentences not pulled from a text, and (2) ungrammatical sentences (often used for one reason or another).
- ★ The same fragment might be used as an example in various contexts, each with different items highlighted. In one context it might illustrate a relative clause; in another, the use of a particular case marker; in the lexicon, the headword of the entry from which it is linked.  
The proto example is enriched in ways that depend on the descriptive context. Highlighting, for example, depends on the claim that the example illustrates.
- ★ The “proto example” might be a collection of partially enriched fragments to be subsequently conflated. For example, a proto example might include a real fragment and a very similar ungrammatical sentence to contrast the difference between these.  
When the proto example includes two or more fragments, the author should be able to indicate the units to be compared or contrasted without being concerned about how this is visually represented.

★ For each point, it must be determined who owns it, that is, who has the right to change it. It is easy to imagine that one linguist records some texts and transcribes them. Another linguist analyzes these, dividing morphemes, adding glosses, and translating them. Yet another enriches them with brackets (to indicate structure), categories. And yet another uses sentences from these enriched texts as examples, but adds functions, indices, . . . . And this is read by one reader who wants to see contrast/comparison represented by conflation, and another reader who wants to see the full fragments with the contrasted parts underlined.

To create examples with diverse enrichments and layouts, linguists need good tools. These must be accessible within hospitable authoring environments. They must be sufficiently powerful to deal with the diversity of types of enrichments and formats. And they must be reasonably easy to learn and use, providing interfaces friendly to the users.