Appendix G

Recommended Analysis Tags

G.1 The Challenge of Tags

Xerox finite-state morphology systems store feature information including part-of-speech category, tense, aspect, mood, person, number, gender, etc. in the form of multicharacter-symbol TAGS such as `+Noun`, `+PresInd`, `+1P`, `+Sg`, and `+Masc`. These tags are in fact just symbols, manipulated exactly like the alphabetic symbols *a*, *b*, and *c*, but they have multicharacter print names. There is nothing magic about `+Noun`, `+PresInd` and `+Sg`; they must all be chosen and declared by the linguist, usually in the Multichar_Symbols section of the **lexc** source file.

Because languages differ so much in the distinctions they make and in the terminology traditionally used to describe those distinctions, imposing rigid rules for defining multicharacter-symbol tags is both impossible and undesirable. However, there is at least *some* benefit to be had if reasonably similar languages adopt, where appropriate, the same tags to mark the same phenomena. If nothing else, this facilitates future maintenance by those who may have to move back and forth from one language to another.

The following guidelines and examples are offered as helpful suggestions, and you are encouraged to follow them, unless of course you have any good reason for doing otherwise. Above all, do not get overly wrought about tag names; remember that

- Real customers will probably never see them;
- Tags do not really mean anything, except as you define them and contrast them with other tags in your system;
- For your own convenience, you want to choose tag names that are informative and yet are not too long to type; and
- It is absolutely trivial to change tag names cosmetically, using Replace Rules.

G.2 Some General Principles for Tag-Name Choice

Prime Directive: Use morphological-analysis tags that are appropriate for indicating the distinctions in the language being analyzed. Do not try to force your language into a descriptive framework that is foreign to it. If there is an established linguistic vocabulary for describing your language, consider defining tags that evoke that vocabulary.

Secondary Directive: Where a set of words act the same syntactically, i.e. where a set of words fit into the same syntactic frames, then they should probably be analyzed with the same tags. Where two words act differently in the syntax, they should probably be analyzed with distinct strings of tags. Use tags, and strings of tags, consistently.

Tertiary Directive: Use the tags and tag orders which have already been used in other **Xerox** products, unless this violates the Prime Directive or Secondary Directive.

G.3 Don't Work in Isolation

It is highly recommended that you *not* work totally in isolation when choosing tags and tag orders. Make a detailed plan as early as you can, selecting a preliminary tagset and defining tag orders, and present this plan to your colleagues for discussion and approval before you continue with further development. Formalize your tagset and tag orders as a regular expression in a LEXICAL GRAMMAR, see section 7.3.1, and use this lexical grammar periodically for testing. You will always need to refine your tagnames and tagset as you progress, but some thoughtful planning at the beginning can help make a more transparent and maintainable system.

G.4 Recommended Tags

If you have no good reason to use a different system, we present the following as recommended tags to choose from.

G.4.1 Major Category Tags

```
+Noun           ! noun (house)
+Prop           ! proper noun (John)
+Verb           ! verb
+Adj            ! adjective
+Adv            ! adverb
+Pron           ! pronoun

+Art            ! article (like English the and a)
+Det            ! determiner (like this, that, those)
```

```
+Quant          ! quantifier (like many, some)

+Conj           ! conjunction (and, or, citatbut)
+Prep           ! preposition

+Title          ! monsieur, senhor, Herr (tend to come before
                ! proper nouns)

+Punc           ! punctuation mark

+Command        ! idiomatic orders, usually military:  march,
                ! fire, shoulder arms, attention, at
                ! ease, shut up (often difficult to
                ! distinguish from the more general imperative
                ! verbs)

+Interj         ! interjections

+Onom           ! onomatopoeia (au-au, guau-guau, cocoricó,
                ! cockle-doodle-do, croak, ribbit,
                ! vre ke kex koax koax (what a frog says in Greek)

+Num            ! number (usually alphabetically based)
+Dig            ! digit-based word

+Prt            ! particle, a hard to define class, sometimes used
                ! for question particles, or like French quoi in
                ! Je l'ai même pas touché, quoi.

+Initial        ! Upper-case letter followed by a period, e.g. Q.
                ! as in many American names:  John Q. Citizen

+Let            ! an individual letter appearing by itself, e.g. q

+For            ! a common foreign word, e.g. the in a French system
```

There are good arguments for placing the major category tag either at the very beginning of the tag string or at the very end. The convention of putting them at the beginning has been used for Spanish, Portuguese, Dutch, Italian, Hungarian, Czech and other systems at **Xerox**.

G.4.2 Subdivisions useful for nouns

```
        ! magnitude

+Aug            ! augmentative
+Dim            ! diminutive

        ! gender
```

```
+Masc
+Fem
+Neut
+MF             ! masculine or feminine, within a system
                !  where most nouns are one or the other
+MN             ! masculine or neuter
+FN             ! feminine or neuter
```

A classifying language will require many more gender-like subdivisions—Dutch required a much more complex gender-marking scheme to accommodate subtle Flemish distinctions (which have mostly been lost in Dutch proper). When in doubt, it is usually better to make distinctions; it is always easy to collapse them later.

In some languages, the following "classes" may also be significant syntactic distinctions.

```
+Hum            ! human
+Anim           ! animate

etc.

        ! case

+Nom            ! nominative case
+Gen            ! genitive
+Acc            ! accusative
+Abl            ! ablative (from)
+Ela            ! elative (out of)
+All            ! allative (to)
+Ess            ! essive (as)
+Par            ! partitive (part of)
+Com            ! comitative (with)
+Abe            ! abessive (without)
+Ine            ! inessive (in)
+Ill            ! illative (into)
+Ins            ! instructive (by)
+Ade            ! adessive (on)
+Tra            ! translative (change to)

        ! number

+Sg             ! singular
+Dual
+Tri
+Pauc           ! paucal (''a few'')
+Pl             ! plural
+SP             ! the word (like English sheep) can be
                !  singular or plural (where most nouns in
                !  the language are either one or the other)
```

Noun tags have typically been ordered in the following way:

```
+Noun+Fem+Sg

+Noun+Dim+Masc+Pl

+Noun+Nom+Masc+Sg
```

If, in the grammatical tradition of your language, the features of a noun are rattled off in a particular order, e.g. “Masculine-Singular Nominative”, then you might best reflect that order in the tag sequences.

G.4.3 Verb Distinctions

In the Romance languages treated so far, tense, aspect and mood were combined into single tags reflecting the realities of the verbal paradigms.

```
+Inf            ! infinitive

+PresInd        ! present indicative
+PresSubj       ! present subjunctive

+ImpInd         ! imperfect indicative
+ImpSubj        ! imperfect subjunctive

+FutInd         ! future indicative
+FutSubj        ! future subjunctive

+Cond           ! conditional
+Impv           ! imperative

+Perf           ! past perfect
+Pluperf        ! pluperfect

+Gerund
+PastPart
```

Depending on your language, you may require other distinctions. You may want to distinguish gerunds and present participles. In that case, we suggest

```
+PresPart
```

Portuguese has inflected infinitives, and so it needs a tag to distinguish these from bare infinitives:

```
+InfFlex
```

In other languages, it might make more sense to use two tags, e.g. `+Pres+Ind` or `+Pres+Subj`, separating the tense and mood, if that better fits the morphotactic system of the language.

```
        ! A reasonable Mood system

+Ind            ! indicative
+Subj           ! subjunctive
+Junc           ! junctive
+Juss           ! jussive
+Opt            ! optative

etc.
```

You may also need individual aspect tags in addition to or instead of tense tags.

```
+Perf           ! perfect
+Imp            ! imperfect

etc.
```

You may also need a system of voice tags:

```
+Act            ! active
+Pass           ! passive
+Middle         ! middle
+Caus           ! causative

etc.

        ! Person tags

+1P             ! first person
+2P             ! second person
+3P             ! third person
```

In some languages, you will need to distinguish between “inclusive we” and “exclusive we” (try `+1P+Pl+Incl` vs. `+1P+Pl+Excl`), third-person animate vs. third-person inanimate (`+3P+Sg+Anim` vs. `+3P+Sg+Inanim`), etc. Yet other languages will require marking `+Neg`, `+Pos`, focus, and other distinctions right in the verb itself. Again, the goal is to choose and consistently use tags that best express the significant morphological/syntactical distinctions in your language. It is impossible and undesirable to force all languages into a rigid system—you will have to decide what is significant and devise the most beautiful tag coding that you can.

G.4.4 Proper Name Distinctions

A language typically has many kinds of proper names, and the distinctions can often be syntactically significant. Given names may typically come before family names, or vice versa, and names for countries, cities and geographical areas may require different syntactic structures.

Unless more distinctions are really needed for the syntax, the following proper-noun subtypes are generally recommended:

Four primary distinctions:

```
+Giv            ! +Prop+Giv for given or first names
+Fam            ! +Prop+Fam     family names
+Place          ! +Prop+Place   any geographical/political
                !                   name
+Org            ! +Prop+Org     any kind of organization,
                !                   e.g. Xerox
```

with a fifth distinction for anything left over.

```
+Misc            ! +Prop+Misc
```

Within the places, there are five standard sub-sub-divisions:

```
+Place
        +Country        ! +Prop+Place+Country   countries
        +City           ! +Prop+Place+City      cities
        +Continent      ! +Prop+Place+Continent continents
        +Region         ! +Prop+Place+Region
                        ! departments, sub-states, provinces;
                        !     especially in the countries
                        !     where the language is spoken
        +Usastate       ! +Prop+Place+Usastate
                        ! states of the USA
```

with a sixth miscellaneous marker for anything left over.

```
         +Misc
```

Depending on your language, it may be necessary to mark gender, case and number, e.g. `Asie+Prop+Place+Continent+Fem+Sg`. As usual, the goal is to choose tags that reflect real distinctions in your language and which may be necessary for a subsequent task such as tagging or parsing.

G.4.5 Pronoun Distinctions

Pronoun systems are very language-specific. Case tags are often useful for distinguishing different classes of pronouns.

```
+Nom            ! nominative
+Dat            ! dative
+Acc            ! accusative
```

Other distinctions, like possessive, may be marked +Gen or +Poss

```
+Gen            ! e.g. +Pron+Gen in a highly case-marked language
+Poss           ! e.g. +Pron+Poss for possessive prons in English
```

Other distinctions, like reflexive, may be limited to the pronoun system.

```
+Rel           ! relative
+Refl          ! reflexive
+Interrog      ! interrogative
+Dem           ! demonstrative

+PrepObj       ! e.g. +Pron+PrepObj  a form of pronoun that
               ! comes only after a preposition
```

Pronouns can also be distinguished by person and number, parallel to verbs.

```
+Pron+Nom+3P+Masc+Sg
```

G.4.6 Adjective Distinctions

The following distinctions have proved useful:

```
+Comp           ! comparative
+Sup            ! superlative
```

Instead of +Sup, one linguist felt obliged to use +Int (for intensive). Augmentative/diminutive, gender and number tags are also appropriate in many languages.

```
+Adj+Masc+Pl
+Adj+Sup+Fem+Sg
+Adj+Dim+Masc+Sg
```

G.4.7 Number-Word Distinctions

Strings representing numbers come in many forms. The following distinctions have proved useful in many languages:

```
+Num           ! for alphabetic strings representing numbers
+Dig           ! for digit-based string
+Rom           ! for Roman numerals
```

Subtypes include

```
+Num+Card       ! for cardinal numbers one and two
+Num+Ord        ! for ordinal numbers first and second
+Dig+Card       ! for 1, 2, 3
+Dig+Ord        ! for 1st, 2nd, 3rd
+Rom+Card       ! for I, II, III, IV
+Rom+Ord        ! for Ie, IIe, IIIe
```

The main-category tag +Num is typically used for strings of alphabetic characters, e.g. *first* might analyze as +Num+Ord (though it may not be distinct from adjectives, in which case it should be marked like any other adjective). Number strings based on strings of digits, like *23*, are usually given the main-category tag +Dig with one or more qualifying tags.

```
1996+Dig+Year
1996

23%+Dig+Percent
23%

23%+Dig+Degree
23
```

deg

```
3.+Dig+Item     ! enumeration of items in a list
3.

3)+Dig+Item
3)

2nd+Dig+Ord
2nd

02/07/95+Dig+Date
02/07/95

555-3496+Dig+Tel
555-3496

$24+Dig+Curr            ! currency expressions
$24

12h24+Dig+Time          ! French format, 24-hour times
12h24
```

```
548-04-1578+Dig+ID      ! USA Social Security number
548-04-1578

38240+Dig+Post          ! a postal code
38240
```

Cardinal numbers are very often marked for gender and number and act exactly like adjectives. In fact, it might be best simply to treat them as adjectives.

G.4.8 Punctuation

The punctuation tag (+Punc) is intended for analyzing input words that consist of a single punctuation symbol, including periods, commas, exclamation marks, etc. They do show up isolated in texts, and some tokenizers will strip off punctuation from words and feed the punctuation tokens separately to the morphological analyzer.

It was once thought useful to distinguish between beginning, middle and ending punctuation, where the positions refer to individual words.

```
+Beg         ! e.g. in English, the left single quote ‘,
             !   the left parenthesis (and other
             !   punctuation marks that typically come
             !   at the beginning of a word
+Mid         ! e.g. in English, underscore, hyphen and
             !   slash, as in  Multichar_Symbols,
             !   twenty-five, and and/or
+End         ! e.g. in English, the right single quote ’,
             !   the right parenthesis, etc.
```

This classification has been problematic. Of course, all the linguist can do is indicate the most likely positioning of the punctuation mark relative to a word.

Taggers seems to need different information, in particular they need to know if a punctuation mark typically ends a whole sentence. In the end, no one really knows how best to analyze isolated punctuation marks within the context of a morphological analyzer. Consult with colleagues and any customers to determine the most useful encoding.

As a purely practical matter it is usually best to handle separated punctuation marks in a separate lexicon and union it with the other lexicons at the end to form the final lexical transducer.

If you need to insert delimiters in your lexical strings, which might include compound boundaries, the boundary between prefixes and the root, etc., use multicharacter symbols for this purpose. Do not use originary punctuation symbols as delimiters in lexical strings.

G.4.9 Article/Determiner Distinctions

```
+Def
+Indef
```

Gender, number, case, etc. tags may also be appropriate.

G.4.10 Adverbs

In non-technical school grammars, all the little words that no one knows what to do with are lumped into a category of Adverbs, a very dangerous pseudo-class. Avoid that tendency, trying to make syntactically real distinctions where appropriate. Some adverbs need only the +Adv tag

```
well+Adv
well
```

Others are obviously, and fairly productively, derived from adjectives, and this is shown by convention with tag strings like the following, wherein ^DB is a multi-character symbols representing a derivation boundary.

```
claro+Adj^DB+Adv
claramente

claro+Adj+Sup^DB+Adv
clarissimamente
```

The secondary tags +Dim, +Aug, +Interrog, +Neg and +Abbr have also been used with +Adv to distinguish various subtypes.

G.4.11 Title Distinctions

Titles may be distinguished by gender, number and other subtags.

```
+Title+Addr     ! for Mr., Mrs., Dr.
+Title+Post     ! for titles that come after a name, such as
                !       Jr. and Esq.
```

G.5 Miscellaneous Problematic Leftovers

```
+Emph        ! to emphasize written words, Dutch can
             ! just add acute accents to words that
             ! are normally written without accents---
             ! the accented forms can be distinguished
             ! by adding an extra +Emph tag on the end
             ! of the normal tags
```

```
+Abbr       ! for abbreviations, e.g.  ‘‘IBM’’ might be
            ! analyzed as ‘‘IBM+Prop+Org+Sg+Abbr’’.
            ! use +Abbr only on the end of other legal
            ! tag strings, not as a principal part of speech

+Pop        ! used on the lexical side (on the end of
            ! a legal string of tags) to signal that
            ! the surface form is popular or informal

+NP         ! a problematic tag used in Portuguese to
            ! show that a proper name acts as a full
            ! noun phrase and does not need (or even
            ! resists) the addition of an article

+Meas       ! a problematic tag used to mark various
            ! measure terms, e.g. km, cm, k (for kilo),
            ! etc.  These might be better marked
            ! +Noun...+Abbr with other appropriate
            ! tags between +Noun and +Abbr reflecting
            ! how the words behave syntactically; if
            ! +Meas is used, it can be subdivided into
            ! +Meas+Distance, +Meas+Size, +Meas+Weight,
            ! +Meas+Temp, etc.

+WordPart   ! when your tokenizer is out of sync with
            ! your morphology, it may break up
            ! multi-word expressions and feed the
            ! analyzer strings of characters that are
            ! not valid words by themselves.  Where
            ! these become a problem (where they show
            ! up a lot during testing), you might want
            ! to enter the worst cases in your
            ! dictionary as +WordPart.  Examples
            ! include the fro in English to and fro
```